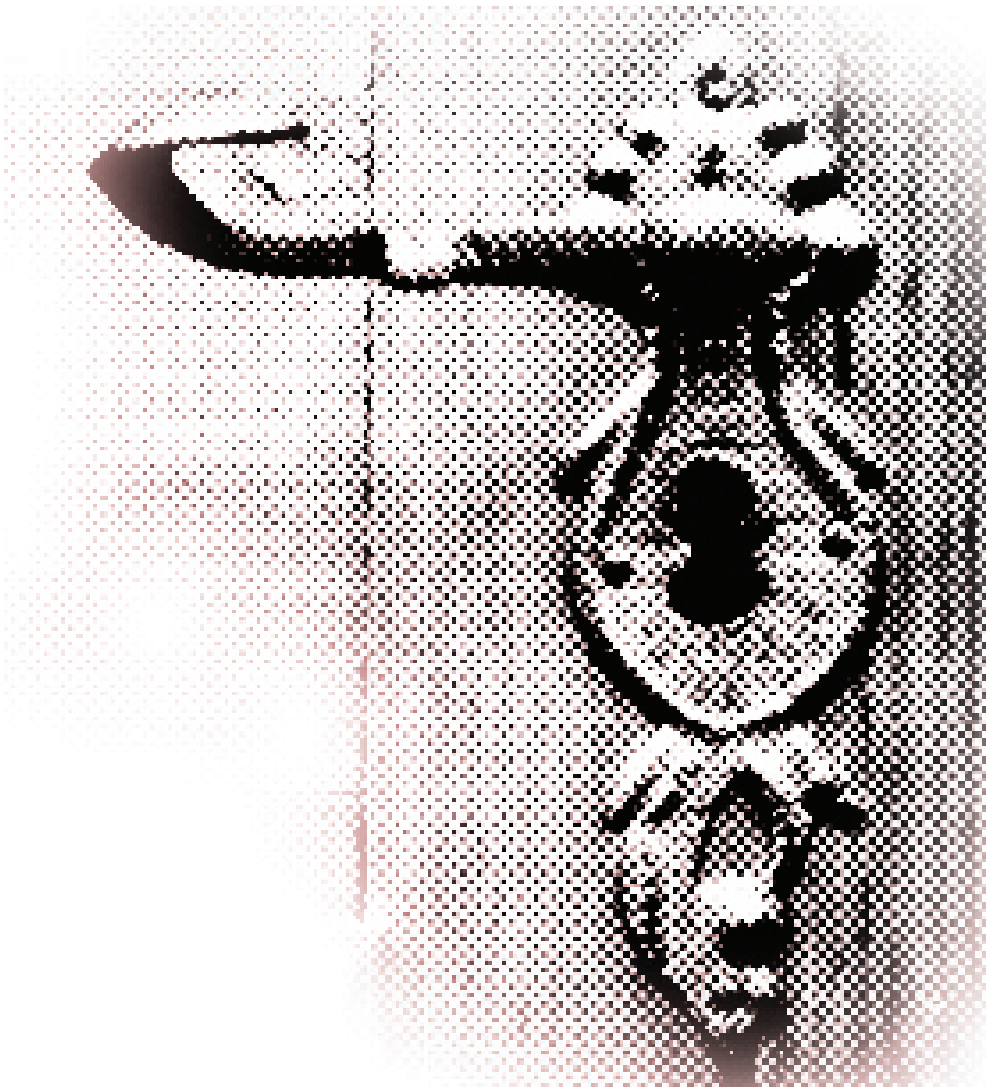
Unlocking the genetics of coeliac disease

Trynka, Malgorzata Barbara

Unlocking the genetics of coeliac disease

Małgorzata Barbara Trynka



Małgorzata, Barbara (Gosia) Trynka

Unlocking the genetics of coeliac disease

Thesis, University of Groningen, with summary in Dutch and Polish

Printing of this thesis was financially supported by Rijksuniversiteit Groningen, University Medical Center Groningen, Groningen University for Drug Exploration (GUIDE) and Celiac Disease Consortium.

Cover design and layout by Claudia Marcela Gonzalez (argo1983@gmail.com)
Printed by EIKON PLUS (Kraków, Poland)

ISBN: 83-60391-65-3

rijksuniversiteit groningen

umcg

GUIDE

cdc
celiac disease consortium

RIJKSUNIVERSITEIT GRONINGEN

Unlocking the genetics of coeliac disease

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
woensdag 14 december 2011
om 12.45 uur

door

Małgorzata Barbara Trynka

geboren op 13 juli 1983

te Kraków, Polen

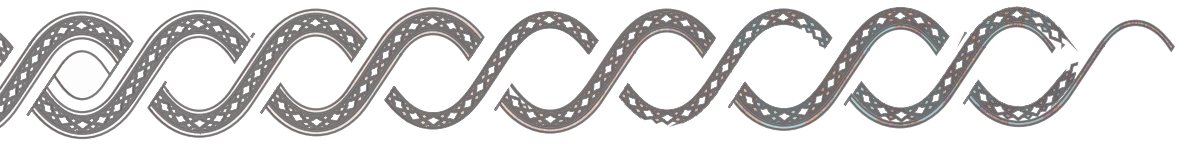
Promotor: Prof. dr. C. Wijmenga

Beoordelingscommissie: Prof. dr. M.P. van den Berg
Prof. dr. G.H. Koppelman
Prof. dr. ir. R.A. Ophoff
Prof. dr. B.H.R. Wolffenbuttel

Dla moich rodziców i ku pamięci babci Pelagii

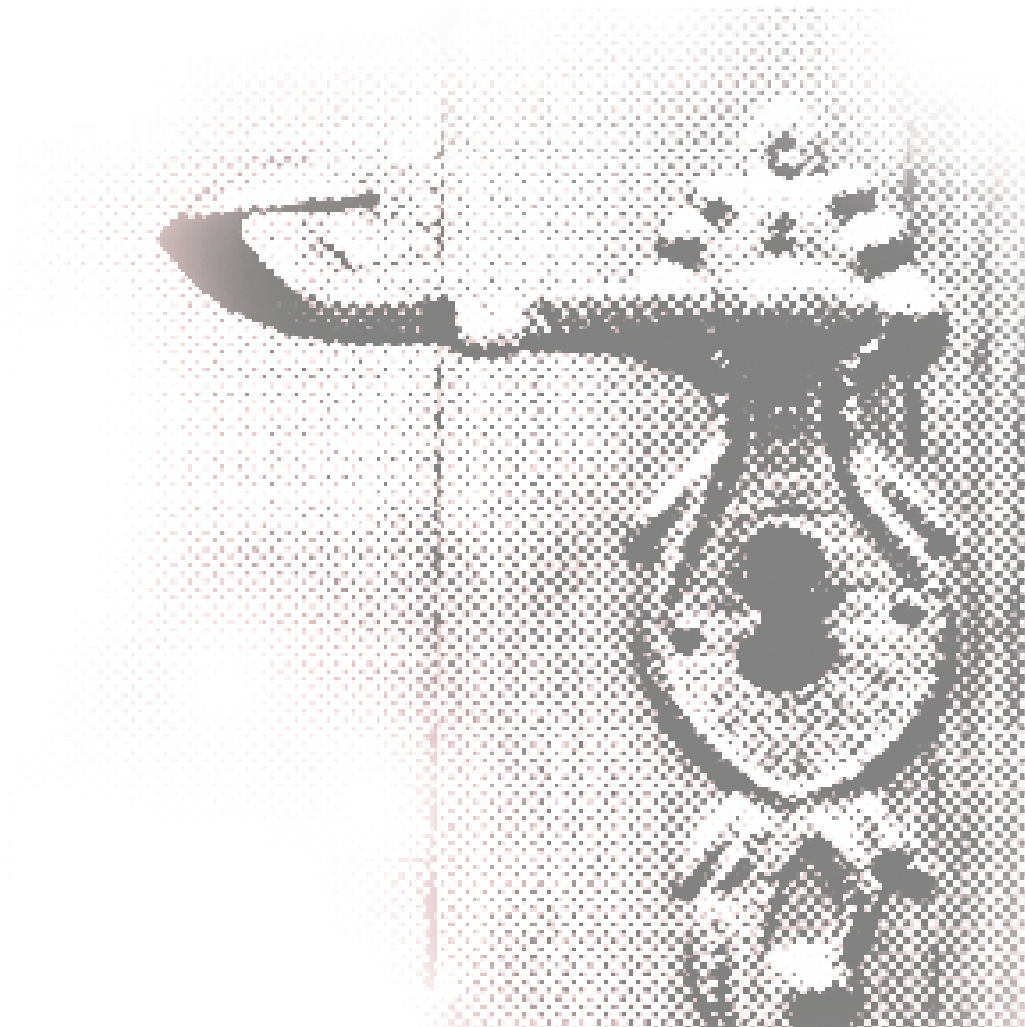
To my parents and in memory of my grandmother, Pelagia

CONTENTS



Preface and outline of the thesis

Chapter 1



Preface

Genetics is one of the most dynamcally evolving fields of science in the past decade. Huntington's disease was the first disease for which a gene was mapped using DNA polymorphisms. This took place in 1983 and the following years yielded rapid advances in discovering genes for rare disorders. However, unlike rare Mendelian diseases, common diseases, such as coeliac disease, type 1 diabetes, inflammatory bowel disease or rheumatoid arthritis, are developed as a consequence of the interplay between multiple genes and one or more environmental triggers. But, with only a few exceptions, until a few years ago the gene mapping for these complex diseases had been rather disappointing. The HapMap release of a Catalogue of Common Human Genetic Variation in 2005[1] and the technical developments in large-scale genotyping platforms led to tremendous advances in gene mapping for complex diseases. The HapMap project provided the first haplotype map of the human genome and showed that most of the common single nucleotide polymorphisms (SNPs) are highly correlated and form block-like structures of linkage disequilibrium (LD). This finding was of great importance as it allowed the testing of only a limited number of SNPs and then, based on their genotypes, the majority of untyped variation could be predicted. Over the past six years, genome-wide association studies (GWAS; Figure 1) that assay hundreds of thousands of SNPs in thousands of affected and healthy individuals have convincingly identified hundreds of associations with dozens of traits and diseases[2].

Coeliac disease is a good example; it is a complex, immune disease and one of the most common food intolerance disorders (Figure 2). It affects around 1% of Western populations but its worldwide prevalence varies (Figure 3). It has a broad spectrum of manifestations[3] and therefore remains largely under-diagnosed (Box 1). The best understood genetic components for coeliac disease are the HLA molecules. The first links between coeliac disease and HLA were reported over 30 years ago[4], and were further characterized in 1989, pointing towards the involvement of particular HLA molecules, *HLA-DQ2* and *HLA-DQ8*[5]. Now, over 20 years later, we have contributed to the identification of an additional 39 non-HLA genes that confer susceptibility to coeliac disease development and have advanced our knowledge about the disease biology.

This thesis had two principal aims: (i) to identify new coeliac disease loci to add to the eight loci known when my thesis work started, and (ii) to further interpret the coeliac disease associations by fine-mapping and cross-ethnic approaches. **Part 1** describes how we identified novel coeliac disease genes using three complementary approaches: (1) by deep replication of the first GWAS for coeliac disease by following up some 500 lower ranked GWAS SNPs, (2) by cross-disease replication, taking loci identified in rheumatoid arthritis and testing them for association in coeliac disease and vice versa, and (3) by performing a large GWAS across four different European populations. **Part 2** describes our fine-mapping efforts to narrow down coeliac disease association signals to smaller genetic intervals, possibly harbouring one or more causal variants. For this, we performed dense, sequencing-based genotyping at all the non-HLA loci and later cross-ethnic gene mapping and replication of the loci in a northern Indian population. **Part 3** discusses our results in the context of coeliac disease and, more broadly, the impact of these studies for other complex disorders. We also outline some future research and clinical perspectives.

Outline of the thesis

Part 1 | Deep Replication and Gene Discovery

In **Chapter 2** I describe the association of two new risk loci for coeliac disease, TNFAIP3 and *REL*, which pointed towards the involvement of the nuclear factor kappa B (NF-κB) pathways. The two regions were identified via deep replication of coeliac disease GWAS results. We selected 458 SNPs that were moderately associated in our first coeliac disease GWAS, which was performed in 2007 on 778 UK cases and 1,422 matched controls, and genotyped these in an extended replication cohort comprising UK, Dutch, Irish and Italian samples.

In **Chapter 3** I describe the genetic background shared between coeliac disease and rheumatoid arthritis. Coeliac disease has a strong autoimmune component and often co-occurs in patients or families with other immune-related diseases, such as type 1 diabetes, inflammatory bowel disease or rheumatoid arthritis. There is a substantial overlap between the genomic regions associated to immune diseases, including coeliac disease and rheumatoid arthritis. We therefore tested if genes associated in one disease can be replicated in the other and vice versa.

In **Chapter 4** I describe how we established association of 13 new loci via a second genome-wide association study of 4,533 coeliac cases and 10,750 controls across four different European populations. We pointed out that over 50% of risk alleles influence gene expression, indicating a potential pathogenic mechanism. I describe how our genetic findings translate to an impaired adaptive and innate immune system, the potential role of viral infection, and the role the thymus plays in coeliac disease pathogenesis.

Part 2 | Dissecting the Genetic Architecture at Coeliac Disease Loci

In **Chapter 5** I review the genetic progress for coeliac disease, its shared background with other diseases and the gene pathways emerging from our GWAS findings.

In **Chapter 6** I describe the use of dense, re-sequencing-based approaches to fine-map coeliac disease loci. I describe the results of genotyping approximately 200,000 markers present on the Immunochip in some 24,000 samples from six countries. This study led to novel regions associated with coeliac disease, bringing the total number of non-HLA genes to 39. For one-third of these coeliac disease regions, we successfully narrowed down the association signal to a small genetic interval. The fine-mapped markers were often located in gene regulatory regions, either 3′ or 5′, suggesting the causative mechanism underlying the disease-alleles will alter gene expression.

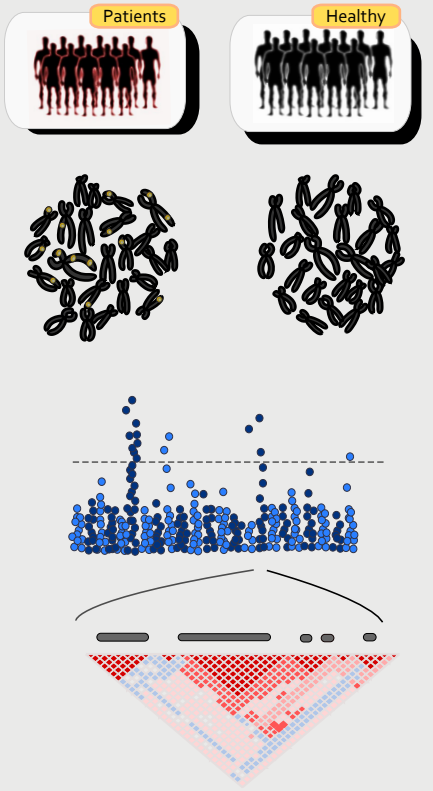
In **Chapter 7** I describe the replication of coeliac disease loci established in Europeans, in a northern Indian population. At the replicated loci, I outline the differences in long-range, inter-marker linkage disequilibrium structure between Europeans and northern Indians. I also point out how, at some loci, the association patterns are mis-localized between the two ethnic groups.

Part 3 | Results in the Context of Coeliac Disease and the Broader Perspective

In **Chapter 8** I place the work described in this thesis in a broader perspective. I discuss the findings described in this thesis and the future research and clinical perspectives for coeliac disease and the field of complex genetics.

Age of onset	from infancy to adulthood
Clinical manifestations	classical: diarrhoea, abdominal distension, abdominal pain, anorexia, flatulence, failure to thrive, muscle wasting, vomiting, weight loss atypical: dermatitis herpetiformis, anaemia, arthritis, osteoporosis, neurological symptoms (e.g. cerebellar ataxia), chronic fatigue, epilepsy, short stature, hepatic steatosis
Diagnosis	histological abnormalities of the small intestine based on a biopsy sample (villous atrophy, crypt hyperplasia), serological tests for elevated tissue transglutaminase (TGA) and endomysial antibodies (EMA), HLA genotyping
Only possible treatment	strict adherence to a life-long gluten-free diet
Consequences	Untreated coeliac disease patients have a higher risk of developing other autoimmune diseases, refractory coeliac disease and enteropathy-associated T-cell lymphoma and also have an increased mortality rate[5].

Box 1 | Characteristics of coeliac disease, its clinical manifestation and diagnosis



Stage 1 | In the first stage, large numbers of patients and healthy controls are recruited, often several hundreds or thousands. The cases and controls need to be ethnically matched to avoid spurious association results.

Stage 2 | Next, whole-genome DNA is isolated and hybridized on chips covering between 300,000 markers (single nucleotide polymorphisms, SNPs), to 2.5 million. SNP alleles are labelled with two fluorescent dyes. The genotypes for each sample and each SNP on the GWAS platform are inferred from the fluorescent signals.

Stage 3 | Allele frequencies for whole-genome markers are calculated and compared between the cases and controls. SNPs with significantly ($<5x10^{-8}$) different frequencies are associated with disease. The results are often represented as a so-called "Manhattan plot", in which the x-axis represents the chromosomes depicted in different colours, and the y-axis shows the significance of the association signal. To ensure meaningful associations are detected, several quality control steps, for samples and SNPs, are performed, e.g. excluding samples with a low call rate (<95%), a population stratification analysis followed by removal of outlying samples, excluding SNPs with low genotype calls (<95%) or those deviating from Hardy-Weinberg equilibrium (p>0.0001).

Stage 4 | The principle of GWAS is based on the linkage disequilibrium (LD) existing between SNPs. The identified disease associations are therefore observed for the SNPs correlating with the real disease variant, rather than being disease-causing themselves. Hence, it is important to interpret the identified signals cautiously, as the associated loci may harbour many plausible disease genes and will need functional follow-up studies to prove causality.

Figure 1 The concept of case-control genome-wide association studies.

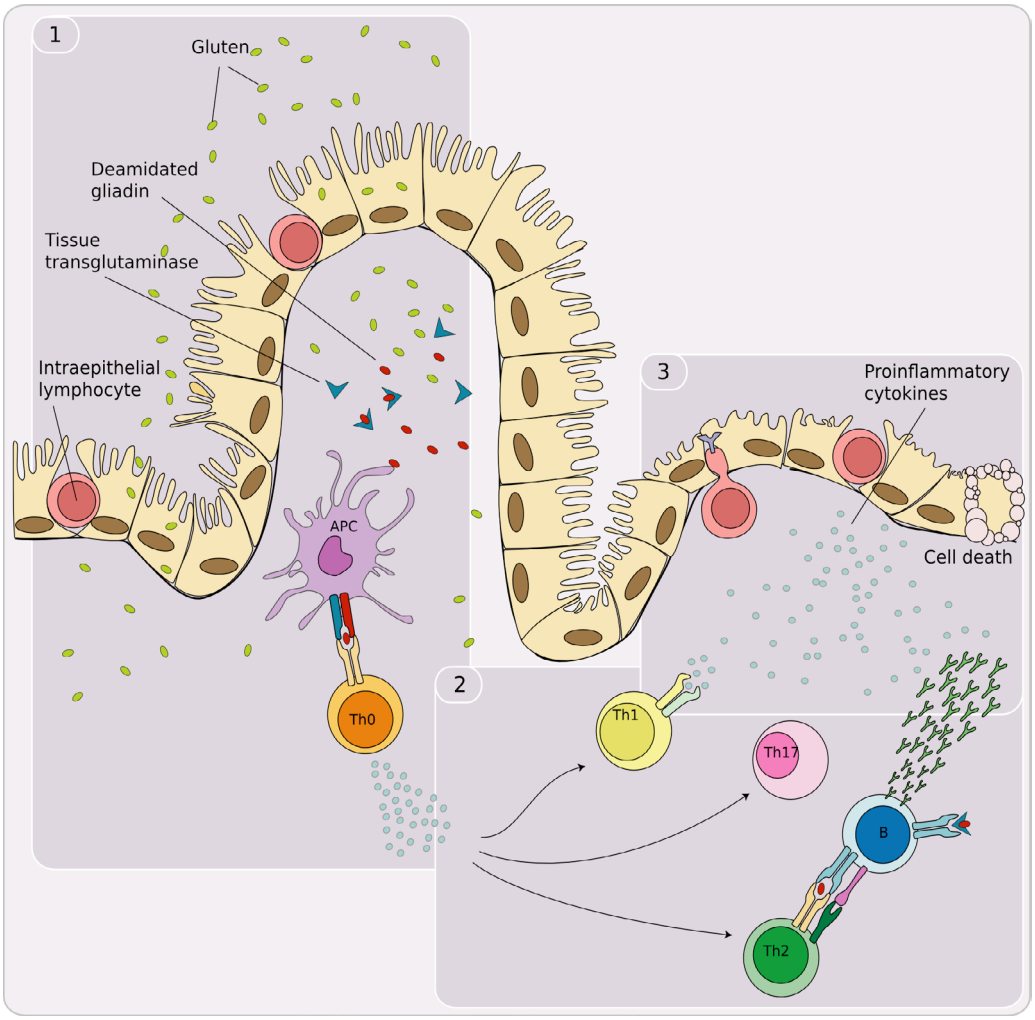
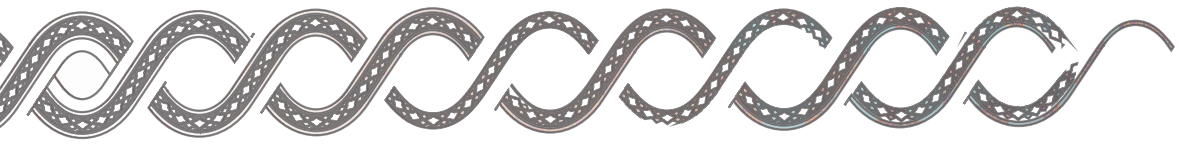


Figure 2. Pathogenesis of coeliac disease
The main mediator of coeliac disease is gluten, a protein present in a wide range of dietary products. Gluten peptides pass through the epithelial barrier of the small intestine into the lamina propria, where they undergo enzymatic modification by tissue transglutaminase (1). This process is called deamidation and leads to an increased immunogenicity of gluten peptides. In the lamina propria, the gluten peptides are 'recognized' as foreign antigens and presented by particular HLA molecules on antigen-presenting cells (APCs) (1). This triggers a cascade of innate and adaptive immune responses and leads to the production of antibodies against gliadin - i.e. anti-endomysium and anti-tissue transglutaminase antibodies - as well as to the production of pro-inflammatory cytokines (2). This inflammatory response results in the destruction of the intestinal epithelium and mucosa, and to lymphocytic infiltration in the proximal part of the small bowel (3). This tissue remodelling eventually causes flattening of the intestinal mucosa, villous atrophy and crypt hyperplasia.

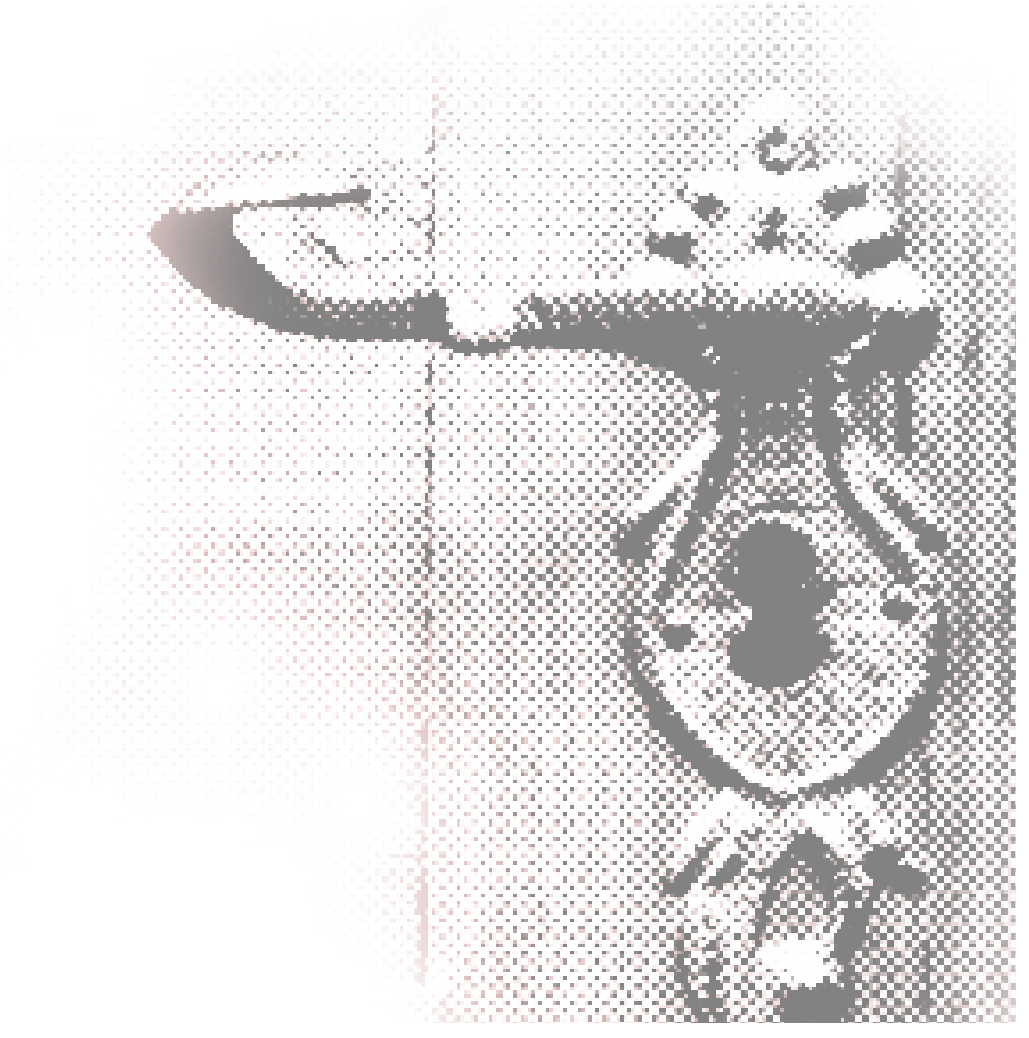


Coeliac associated risk variants in *TNFAIP3* and *REL* implicate altered NF-kappaB signalling

Gut. 2009 Aug;58(8):1078-83

Gosia Trynka*, Alexandra Zhernakova*, Jihane Romanos, Lude Franke, Karen Hunt, Graham Turner, Marcel Bruinenberg, Graham A Heap, Mathieu Platteel, Anthony W Ryan, Carolien de Kovel, Donatella Barisani, Maria Teresa Bardella, Ross McManus, David A van Heel, Cisca Wijmenga

*equal contribution



ABSTRACT

Objective: Our previous coeliac disease genome-wide association study (GWAS) implicated risk variants in the human leucocyte antigen (HLA) region and eight novel risk regions. To identify more coeliac disease loci, we selected 458 single nucleotide polymorphisms (SNPs) that showed more modest association in the GWAS for genotyping and analysis in four independent cohorts.

Design: 458 SNPs were assayed in 1682 cases and 3258 controls from three populations (UK, Irish and Dutch). We combined the results with the original GWAS cohort (767 UK cases and 1422 controls); six SNPs showed associaton with $p<1x10^{-04}$ and where then genotyped in an independent Italian coeliac cohort (538 cases and 593 controls).

Results: We identified two novel coeliac disease risk regions: 6q23.3 (*OLIG3-TNFAIP3*) and 2p16.1 (*REL*), both of which reached genome-wide significance in the combined analysis of all 2987 cases and 5273 controls (rs2327832 $p=1.3 \times 10^{-08}$, and rs842647 $p= 5.2 \times 10^{-07}$) We investigated the expression of these genes in the RNA isolated from biopsies and from whole blood RNA. We did not observe any changes in gene expression, nor in the correlation of genotype with gene expression.

Conclusions: Both *TNFAIP3* (A20, at the protein level) and *REL* are key mediators in the nuclear factor kappa B (NF-kB) inflammatory signalling pathway. For the first time, a role for primary heritable variation in this important biological pathway predisposing to coeliac disease has been identified. Currently, the HLA risk factors and the 10 established non-HLA risk factors explain ~40% of the heritability of coeliac disease.

Coeliac disease is a common intestinal inflammatory disorder, characterised by intolerance to dietary gluten protein from wheat, and related proteins from barley and rye. It is the best understood human leucocyte antigen (HLA) associated disorder. Coeliac disease is rather special because it shares its pathogenesis with other autoimmune diseases (such as type 1 diabetes (T1D) and rheumatoid arthritis (RA)) and with intestinal inflammatory diseases (such as Crohn's disease). Shared genetic risk factors for both coeliac disease and Crohn's disease, as well as for coeliac disease and autoimmunity have been reported.[1–4] To search for genetic risk factors, we recently performed a genome-wide association study (GWAS) in coeliac disease, followed by replication of the 1020 most strongly associated single nucleotide polymorphisms (SNPs) in case–control cohorts from three populations. These studies led to the discovery of eight new non-HLA loci.[1, 5] The results led to three observations: [1]

* Only three of the confirmed loci were presented by SNPs located in the top-100 associated signals in the GWAS, whereas three more associated SNPs ranked below the top-500, and one SNP ranked as low as 1004 in the initial GWAS[2].
* Seven of the eight new loci contained immune-related genes, four of which are cytokines or cytokine receptors (*IL2/IL21*, *IL18RAP*, *IL12A*, and the *CCR1/CCR3* cluster locus).[3]
* Four of the new loci (*IL2-IL21, IL18RAP, CCR3* and *SH2B3*) are shared by other autoimmune and inflammatory disorders. [1-3,5,6]

These three observations prompted us to start a second replication study in which we followed up even lower ranking SNPs from our coeliac disease GWAS, and focused on the involvement of the immune pathways in the pathogenesis of coeliac disease (Fig 1). We therefore enriched our SNP set with SNPs that showed association to coeliac disease in our GWAS and that mapped to the immune-related genes (mostly interleukins and their receptors). In addition, we sought shared autoimmune and inflammatory genes by investigating an overlap between SNPs associated in our coeliac disease GWAS and to either T1D, RA or Crohn's disease in the Wellcome Trust Case Control Consortium (WTCCC) GWAS data. [7]

With such a study design we successfully identified two novel, genome-wide significant, coeliac disease loci: the intergenic region on 6q23.3 located in the proximity of the *TNFAIP3* gene, and 2p16.1, mapping to the second intron of the *REL* gene. Both loci indicate an as yet unrecognised role for the nuclear factor kappa B (NF-κB) signalling pathway in the pathogenesis of coeliac disease.

MATERIALS AND METHODS

Subject DNA

DNA was extracted from whole blood, except for the 1958 cohort control samples, which were lymphoblastoid cell line DNA, and 374 cases and 176 controls from the UK2 collection, which were Oragene saliva DNA. Whole-genome amplified (WGA) blood DNA was used for 194 Irish cases and 18 Dutch cases. Genotype cluster theta values for WGA DNA were similar to blood DNA, for a small fraction of markers the intensity (R) was lower.

Detailed characteristics of UKGWAS, UK2, Irish, Dutch and Italian samples are provided in table 1 and previously published studies.[1 5, 8] Informed consent was obtained from all subjects.

SNP selection

Three groups of SNPs were selected for the genotyping:

- To follow up our GWAS SNPs were selected with p values between p>0.000275 and p<0.004 (indicated as SNP-group 1 in table 2 and supplementary data 1; also indicated as category "WGAtop2000_noWTCCCassoc" in supplemen- tary data 1) (n = 300).

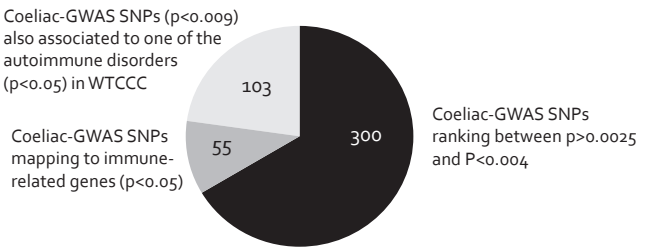


Figure 1 Scheme representing the single nucleotide polymorphisms (SNPs) selected for the second coeliac genomewide association study (GWAS) follow-up study. WTCCC, Wellcome Trust Case Control Consortium.

- SNPs from immune-related genes associated in GWAS with p<0.05 (indicated as SNP-group 2 in table 2 and supplementary data 1; also indicated as category "WGArepli_ ImmuneGenes" in supplementary data 1) (n = 55).
- SNPs within the top-3000 coeliac–GWAS ranking (p<0.009), which showed also association to either type 1 diabetes (T1D), rheumatoid arthritis (RA) or Crohn's disease in the WTCCC study with p<0.05 (n=103). For calculation of association in the WTCCC cohort we used imputed genotypes (WTCCC data accessed on 21 November 2007). T1D, RA and Crohn's cases were compared to the blood donor WTCCC control cohort (n=1500). The 1958 birth cohort was excluded from the WTCCC analysis, as the majority of these samples overlapped with the coeliac disease GWAS control cohort. This SNP category is indicated as SNP-group 3 in table 2 and supplementary data 1; also indicated as category 'WGAtop3000associatedWTCCC" in supplementary data 1).

Golden Gate genotyping

Genotyping of all samples was performed following the manufacturer's protocol. Genotyping data and clustering was performed in BeadStudio. Clustering clouds were manually investigated and adjusted if necessary. Four hundred and fifty-eight SNPs were included in the genotype analysis. Ten SNPs with <95% call rate, because of poor amplification or poor genotype cloud clustering, were excluded. For the top-6 associated SNPs we investigated clustering in subgroups of blood, saliva and lymphoblastoid cell line DNAs. All groups showed similar patterns and comparable theta values. All plates included one duplicate sample to control for plate swaps. Six SNPs were out of Hardy–Weinberg equilibrium (p>0.001) and were excluded from further analysis. In total, 1682 cases and 3258 controls from the three

Table 1 Subjects included in our replication study

	UKGWAS	UK2	Dutch	Irish	Italian	
	Phase I - GWAS	Replication cohort 1			Replication cohort 2	Total
Coeliac cases	767	724	514	444	538	2987
Male/female	213/554	169/555	168/346	142/302	132/406	
Controls	1422	1398	900	960	593	5273
Male/female	720/702	464/934	550/350	284/676	225/368	
TOTAL	2189	2122	1414	1404	1131	8260

populations (replication cohort 1; R1) were successfully genotyped for 442 SNPs.

Genotype concordance

A single control DNA sample was included in each 96-well plate. Genotype concordance for this sample was 99.9% for 45 replicates of 442 SNPs.

Additional genotype quality control

Pairwise comparisons of identity-by-descent were made for all samples (UK2, Irish and Dutch) using PLINK v1.02. We detected the same proportion of first-degree relatives as described by Hunt *et al* and excluded one sample from each pair of first- degree relatives from the entire dataset in the current study.[1] All of the top association findings were in Hardy–Weinberg equilibrium in controls.

Taqman genotyping

Replication cohort 2 (Italian population) was genotyped using TaqMan probes and primers developed by Applied Biosystems, on an ABI 7900HT system (Applied Biosystems, Nieuwerkerk a/ d IJssel, the Netherlands). Genotyping was performed following the manufacturer's specifications. DNA samples were processed in 384-well plates and each plate contained eight negative controls and 16 genotyping controls (four duplicates of four different samples obtained from the Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France).

Genotype statistical analysis

Cochran–Mantel–Haenszel allele count $\chi 2$ association tests were performed using PLINK[9] with four clusters: UKGWAS (Infinium assay), UK2, IRISH and DUTCH (Golden Gate assay) collections. All p values are two-tailed. The Cochran–Mantel–Haenszel allele count χ^2 association test implicated in SPSS v 15 was used for combined analysis of association of the coeliac disease cohort (including R2 (Italian) samples), and for analysis of the combined autoimmune cohort. Using Tarone's statistic in SPSS v 15, we tested for heterogeneity of odds ratios between the different coeliac disease cohorts for the SNPs reported in table 2. The odds ratios differed significantly between cohorts (p=0.007) only for rs1160542. Haplotypes and linkage disequilibrium (LD) blocs were defined and analysed using Haploview v4.1.[10]

Intestinal biopsy expression analysis

The duodenal tissue biopsies from 12 healthy controls, 12 untreated coeliac individuals with Marsh III and 12 treated coeliac individuals with

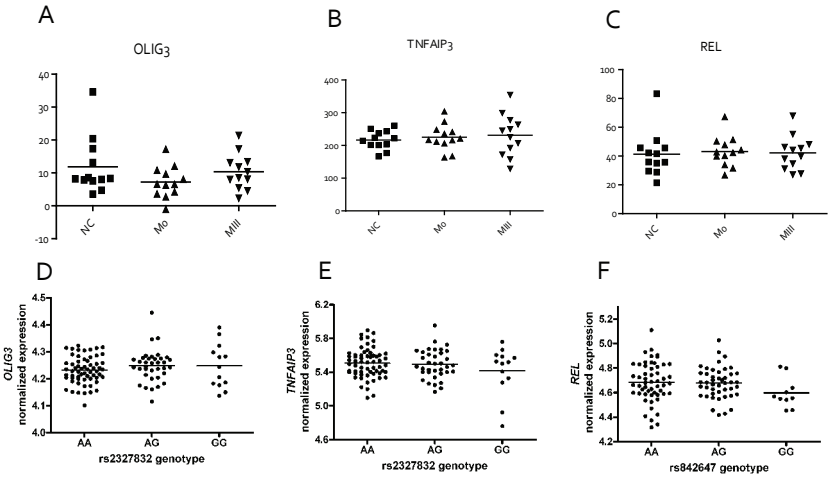


Figure 2 Expression of candidate genes in small intestine tissue from 12 normal healthy controls (NC), 12 treated coeliac disease patients with a normal histology of the small intestine, Marsh 0 (Mo) and 12 untreated coeliac disease patients with villous atrophy, Marsh III (MIII) (A–C). Correlation of the rs2327832 genotype with OLIG3 and TNFAIP3 (D,E) and rs842647 genotype with REL expression (F). Expression levels were determined from the whole blood PAXgene samples from 110 patients with coeliac disease on a gluten-free diet. Bars show group means (A–F).

Table 2 Association of single nucleotide polymorphisms (SNPs) with $p<10^{-4}$ in the replication cohorts, WTCCC data, and in the combined analysis

Genes in proximity	Chr.	SNP	SNP group	A1	A2	P_CD_GWAS	CMH_R1	CMH_R1 +GWAS	OR_R1+GWAS	P value R2	OR R2	CMH GWAS+R1+R2	OR GWAS +R1+R2	P_WTCCC Crohn*	P_WT CCC T1D*	P_WT CCC RA*	CMH _AID**
REL	2	rs842647	2	G	A	0.0274	0.00010	9.26E-06	0.85	0.016	0.79	5.20E-07	0.84	0.8478	0.1157	0.7909	0.002
REL/PEX13	2	rs1729657	1	A	G	0.0029	0.00657	1.00E-04	0.87	0.204	0.89	4.75E-05	0.87	0.7933	0.9226	0.8118	0.001
AFF3 - LONRF2	2	rs1160542	3	G	A	0.0059	0.00360	6.68E-05	1.15	0.048	0.84	3.97E-03	1.10	0.8781	0.0007	0.0037	2.49E-06
CD80 OLIG3 - TNFAIP3	3	rs1599796	2	A	G	0.0041	0.00241	4.07E-05	1.20	0.281	1.12	3.16E-05	1.18	0.1056	0.1505	0.9814	0.112
TCC22D1	6	rs2327832	3	G	A	0.0029	0.00063	6.61E-06	1.21	7.9E-05	1.53	1.31E-08	1.25	0.2438	0.0046	0.0001	3.78E-12
	13	rs9533935	3	C	A	0.0048	0.00594	9.13E-05	1.15	0.835	1.02	2.04E-04	1.13	0.6872	0.0151	0.9735	0.121

SNP group indicates the group of SNPs selected (as indicated in the text and supplementary information online).
R1 (replication cohort 1) included 1648 cases and 3258 controls from three populations.
R2 (replication cohort 2) included 538 cases and 593 controls from Italy.
The cohort used in our genome-wide association study (GWAS) in coeliac disease had 767 cases and 1422 controls (see supplementary information online).
*p Values calculated for the WTCCC cases versus NBS (normal blood sample) controls.
**p Values calculated using Cochran–Mantel–Haenszel in which all four coeliac disease populations were summed and considered as one cohort, and Crohn's disease, RA and T1D cohorts from WTCCC (imputed genotypes) were compared to NBSWTCCC controls.
AID, autoimmune diseases, here includes coeliac disease (all samples included in the current project), and WTCCC cohorts of Crohn's disease, type 1 diabetes (T1D) and rheumatoid arthritis (RA); Chr, chromosome; CMH, Cochran–Mantel–Haenszel
meta-analysis; OR, odds ratio; WTCCC, Wellcome Trust Case Control Consortium.

Marsh 0 were collected in RNAlater (Applied Biosystems/Ambion, Austin, Texas, USA). RNA was extracted using TRIzol (Invitrogen, Carlsbad, California, USA) and glass beads, hybridised to HumanRef-8v2 arrays and analysed as previously described[1]

Quality control of whole blood PAXgene data
The whole blood PAXgene expression data from 110 unique coeliac individuals, who were also genotyped in the UKGWAS, was isolated, hybridized to the Illumina HumanRef-8v2 arrays and analysed as previously described.[11]

RESULTS

In a combined analysis of 2449 cases and 4680 controls (Replication cohort 1 (R1) + GWAS), eight SNPs were associated with $p<1x10^{-04}$. The strongest association signal was observed for rs2327832 in the *OLIG3-TNFAIP3* locus ($pCMH(GWAS+R1)=6.6x10^{-06}$) (see supplementary data 1 for the results of SNPs per population and combined results). Six SNPs with $p<1x10^{-04}$ (combined GWAS+R1 cohort) were further genotyped in an independent cohort from Italy, comprising 538 cases and 593 controls (Replication cohort 2, R2). We did not genotype rs842639 which was strongly correlated with rs842647 (r2>0.95), and rs4851274, which failed the Taqman design. Of these six SNPs, two SNPs were replicated with associations to the same allele (rs2327832 $p=7.93x10^{-5}$, and rs842647 p = 0.015). rs1160542 in the *AFF3-LONRF2* locus, showed a trend of association to the opposite allele (p=0.048). Combining the results from our GWAS and the two replication cohorts revealed two SNPs (rs2327832 in the (*OLIG3/TNFAIP3* locus and rs842647 in the *REL* locus) convincingly associated with coeliac disease with $p=1.3x10^{-08}$ (odds ratio (OR) = 1.25; confidence interval (CI), 1.15 to 1.34) and p = 5.2 x 10-07 (OR = 0.84; CI, 0.78 to 0.90), respectively.

In addition, to search for independent association signals in coeliac disease we performed haplotype analysis with 85 coeliac-GWAS genotyped SNPs encompassing the *OLIG3-TNFAIP3* region (chromosome 6;

bp137851837–138241110, NCBI build36) using the 767 coeliac cases and 1422 controls included in our GWAS study. However, none of the other single SNPs or haplotypes was more strongly associated than rs2327832 (data not shown).

To assess the functional role of the SNPs in two associated regions, we have investigated the available datasets of the genome-wide association studies of gene expression.[12, 13] No significant effect of the rs2327832 SNP on the *OLIG3* or *TNFAIP3*, nor of rs842647 on *REL*, was observed in these datasets. We also compared the RNA expression profiles of the *OLIG3*, *TNFAIP3* and *REL* genes in small-intestine tissue from healthy controls and from treated and untreated coeliac patients. None of the genes showed significant differential expression between the three groups (Fig 2A–C).

We correlated *cis* gene expression in the whole blood samples from 110 patients with coeliac disease for the *OLIG3/TNFAIP3* and rs2327832, as well as *REL* with rs842647 genotypes. No significant effect of genotype on gene expression was observed (Fig 2D–F). We also investigated if any other SNP in the 1 Mb window around each of the two associated SNPs affected gene expression. No *cis* effect was observed after correcting for multiple testing (supplementary data 2).

As one of our aims was to search for shared autoimmune genes, we also combined our results with those from the WTCCC GWAS, in particular the results for RA, T1D and Crohn's disease compared to the WTCCC blood donor control group. Two of the variants were found

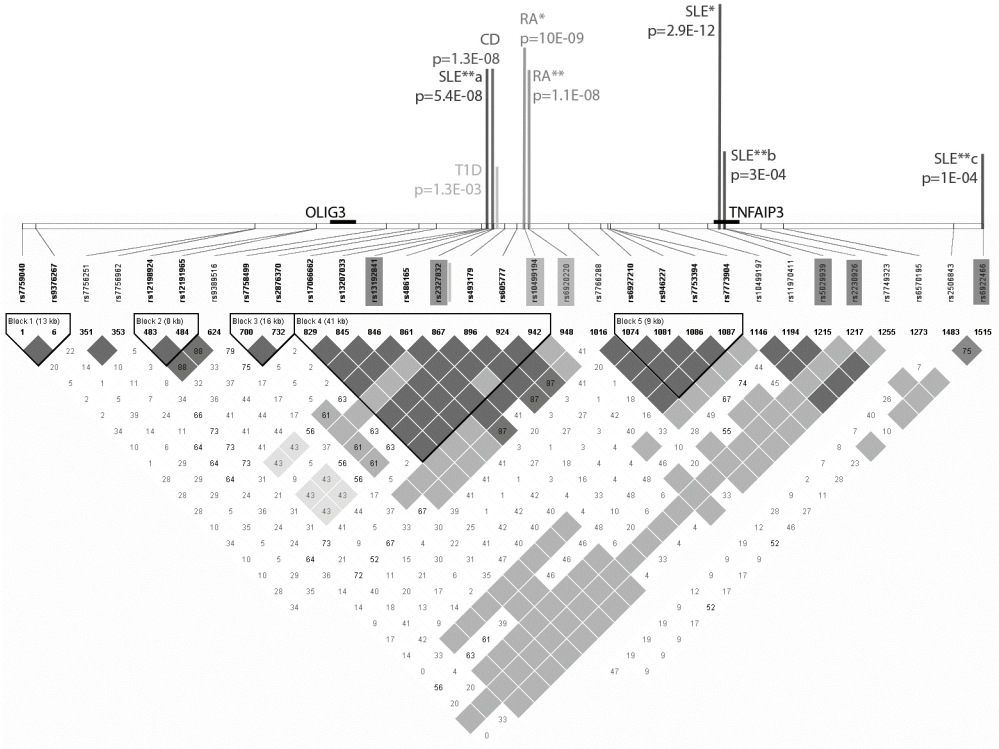


Figure 3 Location of the signals in the four autoimmune diseases associated to the *OLIG3/TNFAIP3* locus. The figure shows the linkage disequilibrium (LD) block between the *OLIG3* and *TNFAIP3* genes (NCBIb36, 137 559 184 bp to 138 486 672 bp) associated with multiple diseases. The plot is based on HapMap CEU data; the D' plot was generated by Haploview. Single nucleotide polymorphims (SNPs) associated per disease are: SLE*a,b,c and SLE** (SNPs reported to be associated with SLE (Graham et al16 and Musone et al,17 respectively); RA* and RA** (SNPs associated with RA in Plenge et al14 and Thomson et al,15 respectively); T1D (SNP associated to T1D in the WTCCC study7); coeliac disease (SNP associated to coeliac disease in the current study). RA, rheumatoid arthritis; SLE, systemic lupus erythymatosus; T1D, type 1 diabetes; WTCCC, Wellcome Trust Case Control Consortium.

to be associated to T1D and RA in the WTCCC GWAS: the *OLIG3/TNFAIP3**rs2327832 SNP and the *AFF3-LONRF2**rs1160542 SNP. All associations in the WTCCC dataset were observed to the same allele as in our coeliac disease analysis. Combining the analysis of all inflammatory diseases (including coeliac disease, T1D, RA and Crohn's disease) rs2327832 in *OLIG3/TNFAIP3* ($p = 3.78 \times 10^{-12}$) showed genome-wide significance to immune-related diseases (table 2). A second SNP, rs1160542 in *AFF3-LONRF2*, showed modest association ($p = 2.49 \times 10^{-06}$). The *AFF3-LONRF2* locus has previously been nominally replicated in T1D in an independent case/control (p=0.02)and a family (p=0.01) datasets,[5] further strengthening a role for this locus in immune-related disorders.

DISCUSSION

In this study we performed an extensive replication of moderately associated genetic variants from the GWAS study in coeliac disease, including variants located in immune-related genes, and potentially shared autoimmune SNPs. The strongest association to both coeliac disease and immune-related diseases was seen for rs2327832, located in a 60 kb block of linkage disequilibrium (LD) between the *OLIG3* and *TNFAIP3* genes. Interestingly, two independent SNPs located in the same LD block have been previously associated to RA;[14, 15] one of the RA associated SNPs (rs6920220) is a perfect proxy (r2 = 1 in CEU HapMap samples) for rs2327832, the SNP found to be associated in this study. Recently, the same region was found to be associated to systemic lupus erythematosus (SLE), another immune-related disease.[16 17] For SLE a second, independent variant in the *TNFAIP3* gene was also found to be associated.[16] This indicates the *TNFAIP3* gene region as a new, shared autoimmune locus. The scheme of association among different autoimmune disorders is presented in Fig 3, indicating the complexity of the association pattern within this locus. Both the same and different variants are associated to various autoimmune traits.

TNFAIP3 is an attractive candidate for both inflammatory and autoimmune pathogenesis. The *TNFAIP3* gene product A20 is required for termination of the NF-κB signal mediated by innate immune receptors via the de-ubiquitylation of several NF-κB signalling factors.[18] Genetic deficiency of A20 in mice leads to persistent activation of NF-κB by toll-like receptors, resulting in multi-organ inflammation, cachexia and neonatal lethality.[19, 20] It has been suggested that loss of A20 breaks down the tolerance of the innate immune system to the commensal intestinal microflora.[21] Although we could not observe an effect of the associated polymorphisms on expression, this does not exclude a role for A20 in coeliac disease; there might well be an effect on the protein level. A20 regulation is rather complex and can be modified by A20-binding proteins such as ABINs or TAX1BP1, as well as by post-translational modifications in A20 protein.[22] We do not exclude that subtle changes in the protein structure could lead to modification of A20 activity and, together with other coeliac-associated risk variants, cause the disease. Further extensive studies, including fine mapping with sequencing as well as functional studies (including those on a protein level), are required to identify the true causal variants.

The second new gene associated with coeliac disease – *REL* – is a component of the NF-κB transcription complex that plays a critical role in promoting immune and inflammatory responses including through the production of pro-inflammatory cytokines. In another study we observed a moderate association of *REL* polymorphisms to ulcerative colitis ($p = 0.001$), another intestinal inflammatory disorder,[3] suggesting that this gene may not be unique to coeliac disease pathogenesis. Association of coeliac disease to both *TNFAIP3* and *REL* points to a role for innate signalling via NF-κB in the pathology of coeliac disease, this is a novel finding and has not been reported before.

In this study we have extended the replication of our GWAS in coeliac disease and searched for genes shared by coeliac disease and other

autoimmune and inflammatory intestinal disorders. We discovered two new loci associated to coeliac disease: *REL* and *OLIG3/TNFAIP3*. The *OLIG3/TNFAIP3* locus can be considered to be a general immune-related locus as it has now been associated to four autoimmune disorders (Fig 3). This supports the recent observation that many disease susceptibility genes contribute to multiple diseases.[23] So far, the pathways associated with coeliac disease have pointed to T cell signalling and multiple cytokine involvement. Our observation that the NF-κB signalling pathway is also important adds a new player to the field.

NF-κB is a transcription complex that plays a key role in regulating the cellular immune response to infections, stress, cytokines and other stimuli. Activation of NF-κB in various inflammatory disorders, including asthma, arthritis and inflammatory bowel disease (IBD) has been described.[24]

Coeliac disease can now be added to the list of complex disorders that show association to the 6q23 region. This strengthens the importance of A20 in controlling inflammation in autoimmune diseases and points to A20 as an attractive candidate for drug targeting, as suggested recently by Coornaert et al.[22]

How the newly discovered genes interplay with the previously identified coeliac loci can only be speculated. On the one hand, activation of the NF-κB complex leads to overexpression of inflammatory cytokines and, together with previously identified cytokines and cytokine receptor genes such as *CCR5, RGS1, IL12A* and *IL18RAP*, this pathway would be important in fine-tuning the immune response. On the other hand, the NF-κB pathway may play an independent role in the innate mechanisms of disease development. Strikingly, genes involved in the innate immune response have recently been associated with various autoimmune diseases, suggesting a role for microbial and viral triggers in disease development. In coeliac disease an increased frequency of rotaviral infections have been observed, suggesting that viral infections may, for example, trigger an innate immune response.[25]
It is interesting that the knockdown studies of A20 in dendritic cells show a shift in the subset of activated T cells, hyperactivation of cytotoxic T lymphocytes and T helper cells, and suppression of regulatory T cells. This shift results from enhanced expression of co-stimulatory signals and proinflammatory cytokines when inhibiting A20[26] and would fit in celiac disease being a Th1 mediated disease.

These two novel loci can be added to the list of eight known, non-HLA, genetic risk factors for coeliac disease that have a smaller risk effect than HLA. Extending the list of common variants that account for coeliac disease will improve the genetic prognosis of patients and may help to predict the likelihood of individuals from the at-risk group developing coeliac disease.

Acknowledgements: We thank J Swift, P Kumar, D P Jewell, S P L Travis, L Dinesen and K Moriarty for collection of UK-GWAS and additional coeliac case samples. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant Goooo934 and the Wellcome Trust grant 068545/Z/02. We thank H van Someren and F Mulder for clinical database management, E Oostrom, R van 't Slot and the genotyping facilities at UMCG and UMC Utrecht (the Netherlands) for technical assistance. We thank C Feighery and J McPartlin for sample collection. Irish control DNA was supplied by the Irish Blood Transfusion Service/ Trinity College Dublin Biobank. We thank all coeliac and control individuals for participating in this study. We thank J Senior for critically reading the manuscript.

Funding: The study was supported by grants from Coeliac UK (to DAvH); the Coeliac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to CW); the European Union (STREP 036383); the Netherlands Organization for Scientific Research (VICI grant 918.66.620 to CW); the Science Foundation Ireland; the Higher Education Authority PRTLI; The Irish Health Research Board; and the Wellcome Trust (GR068094MA Clinician Scientist Fellowship to DAvH; New Blood Fellowship to RMcM).

Competing interests: None.

Ethics approval: Ethics approval was from Oxfordshire REC B or East London and the City REC 1 (UKGWAS, UK2), the Medical Ethical Committee of the University Medical Centre Utrecht (Dutch), the Institutional Ethics Committee

of St James's Hospital (Irish) and the Ethics Committee of the Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena (Italian).

REFERENCES

1. Hunt KA, Zhernakova A, Turner G, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 2008;40:395–402.

2. Zhernakova A, Alizadeh BZ, Bevova M, et al. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am J Hum Genet 2007;81:1284–8.

3. Zhernakova A, Festen EM, Franke L, et al. Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. Am J Hum Genet 2008;82:1202–10.

4. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 2009;10:43–55.

5. van Heel DA, Franke L, Hunt KA, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 2007;39:827–9.

6. Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39:857–64.

7. Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.

8. Romanos J. Six new celiac disease loci replicated in an Italian population confirm association to celiac disease. J Med Genet 2009;46;60–3.

9. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.

10. Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.

11. Heap GA, Trynka G, Jansen RC, et al. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med Genomics 2009;2:1.

12. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. Nat Genet 2007;39:1202–7.

13. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. Nat Genet 2007;39:1217–24.

14. Plenge RM, Cotsapas C, Davies L, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 2007;39:1477–82.

15. Thomson W, Barton A, Ke X, et al. Rheumatoid arthritis association at 6q23. Nat Genet 2007;39:1431–3.

16. Graham RR, Cotsapas C, Davies L, et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat Genet 2008;40:1059–61.

17. Musone SL, Taylor KE, Lu TT, et al. Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nat Genet 2008;40:1062–4.

18. Sun SC. Deubiquitylation and regulation of the immune response. Nat Rev Immunol 2008;8:501–11.

19. Boone DL, Turer EE, Lee EG, et al. The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses. Nat Immunol 2004;5:1052–60.

20. Lee EG, Boone DL, Chai S, et al. Failure to regulate TNF-induced NF-kappaB and cell death responses in A20-deficient mice. Science 2000;289:2350–4.

21. Turer EE, Tavares RM, Mortier E, et al. Homeostatic MyD88-dependent signals cause lethal inflamMation in the absence of A20. J Exp Med 2008;205:451–64.

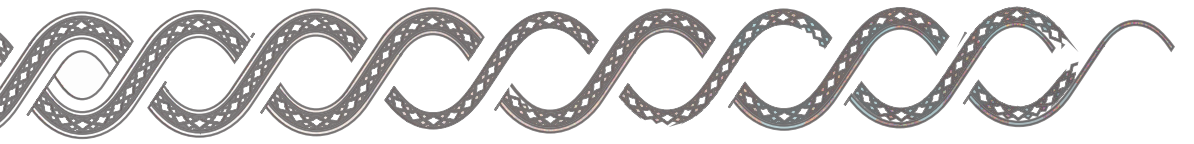
22. Coornaert B, Carpentier I, Beyaert R. A20: Central gatekeeper in inflammation and immunity. J Biol Chem 2009;284:8217–21.

23. Xavier RJ, Rioux JD. Genome-wide association studies: a new window into immune-mediated diseases. Nat Rev Immunol 2008;8:631–43.

24. Perkins ND. Integrating cell-signalling pathways with NF-kappaB and IKK function. Nat Rev Mol Cell Biol 2007;8:49–62.

25. Stene LC, Honeyman MC, Hoffenberg EJ, et al. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am J Gastroenterol 2006;101:2333–40.

26. Song XT, Evel-Kabler K, Shen L, et al. A20 is an antigen presentation attenuator, and its inhibition overcomes regulatory T cell-mediated suppression. Nat Med 2008;14:258–65.

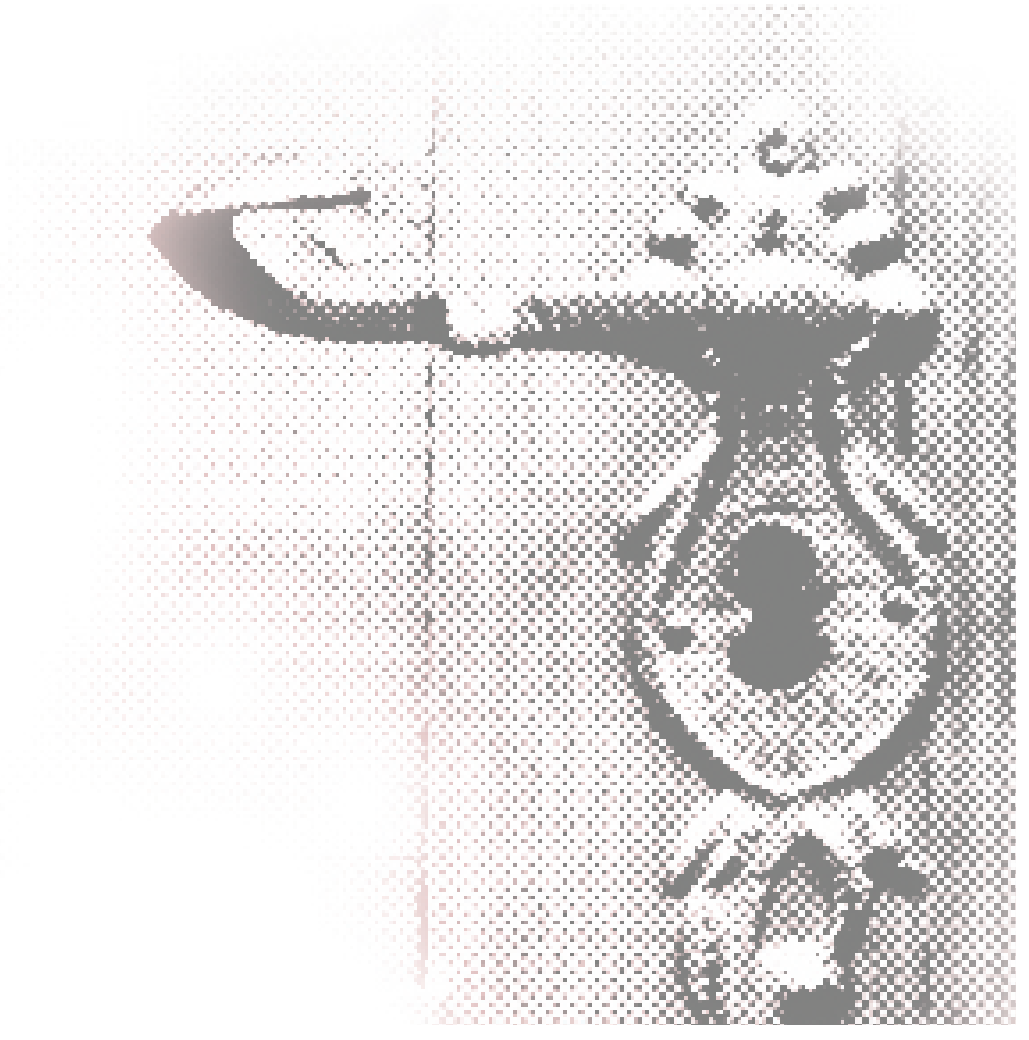


Common and different genetic background for rheumatoid arthritis and coeliac disease

Hum Mol Genet. 2009 Nov 1;18(21):4195-203

Gosia Trynka*, Marieke J H Coenen*, Sandra Heskamp, Barbara Franke, Cleo C van Diemen, Joanna Smolonska, Miek van Leeuwen, Elisabeth Brouwer, Marike H Boezen, Dirkje S Postma, Mathieu Platteel, Pieter Zanen, Jan-Willem W J Lammers, Harry J M Groen, Willem P T M Mali, Chris J Mulder, Greetje J Tack, Wieke H M Verbeek, Victorien M Wolters, Roderick H J Houwen, Luisa M Mearin, David A van Heel, Timothy R D J Radstake, Piet L C M van Riel, Cisca Wijmenga, Pilar Barrera and Alexandra Zhernakova

*equal contribution



ABSTRACT

Recent genome-wide association studies (GWAS) have revealed genetic risk factors in autoimmune and inflammatory disorders. Several of the associated genes and underlying pathways are shared by various autoimmune diseases. Rheumatoid arthritis (RA) and coeliac disease (CD) are two autoimmune disorders which have commonalities in their pathogenesis. We aimed to replicate known RA loci in a Dutch RA population, and to investigate whether the effect of known RA and CD risk factors generalize across the two diseases. We selected all loci associated to either RA or CD in a GWAS and confirmed in an independent cohort, with a combined P-value cut-off $P < 5 \times 10^{-6}$. We genotyped 11 RA and 11 CD loci in 1368 RA patients, 795 CD patients and 1683 Dutch controls. We combined our results in a meta-analysis with UK GWAS on RA (1860 cases; 2938 controls) and CD (767 cases; 1422 controls). In the Dutch RA cohort, the *PTPN22* and *IL2/IL21* variants showed convincing association ($P = 3.4 \times 10^{-12}$ and $P = 2.8 \times 10^{-4}$, respectively). Association of RA with the known CD risk variant in the *SH2B3* was also observed, predominantly in the subgroup of rheumatoid factor-positive RA patients (P = 0.0055). In a meta-analysis of Dutch and UK data sets, shared association with six loci (*TNFAIP3, IL2/IL21, SH2B3, LPP, MMEL1/TNFRSF14* and *PFKFB3/PRKCQ*) was observed in both RA and CD cohorts. We confirmed two known loci and identified four novel ones for shared CD–RA genetic risk. Most of the shared loci further emphasize a role for adaptive and innate immunity in these diseases.

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disorder affecting 1% of the adult population. The disease is characterized by inflammation of the synovial tissue of multiple joints leading to pain, deformities and a reduced quality of life. The aetiology of RA is complex and largely unknown; however, epidemiologic data support a polygenic susceptibility model[1]. Besides this, environmental factors may also play a role in the pathogenesis of RA[2].

RA shares common pathogenic mechanisms with other autoimmune disorders. This is reflected by the co-occurrence of several autoimmune disorders in some patients and families, and the shared genetic background of autoimmune diseases[3,4]. One of the autoimmune disorders showing similar pathogenic mechanisms to those seen in RA is coeliac disease (CD). This is a chronic intestinal inflammatory disorder which develops through an intolerance to gluten, the major dietary protein in wheat, and related proteins from barley and rye[5]. Although the diseases are clearly distinct in their phenotype, common features in RA and CD include the association of the HLA molecules with the diseases, T-cell infiltration in the target organs, the development of disease-specific auto-antibodies and the role of enzymes involved in post-translational modifications in the pathogenesis of the diseases[6]. In addition, several studies show co-occurrence of CD and RA[7–10]. However, there are no well-designed studies to assess the true prevalence of RA and CD co-morbidity.

Recently performed genome-wide association studies (GWAS) have identified 11 loci associated with RA and 10 loci associated with CD, in addition to the already known HLA loci[11–21] (Supplementary Material, Table S1). These GWAS showed that, besides HLA, two chromosomal regions are shared between RA and CD: a region on chromosome 4q27 (including the genes *IL2* and *IL21*) and the 6q23.3 locus containing the *OLIG3* and *TNFAIP3* genes[22]. Interestingly, the *IL2/IL21* locus has also been associated with psoriatic arthritis, Grave's disease and type 1 diabetes, whereas the *OLIG3/TNFAIP3* locus has also been associated to systemic lupus erythematosus and type 1 diabetes[13,23–27]. These results strongly suggest that overlapping genetic mechanisms underlie the development of multiple autoimmune disorders[4]. We therefore hypothesize that other susceptibility genes identified for RA might also contribute to the development of CD and vice versa. In this study, we investigated a total of 22 SNPs associated with either RA (11 SNPs) or CD (11 SNPs) by association testing in Dutch RA and CD cohorts and by meta-analyses including well-comparable data sets from two earlier GWAS on RA and CD.

RESULTS

In this study, we performed an association analysis with all SNPs showing replicated association to either RA or CD with a P-value cut-off of 5×10^{-6}. In total, 22 SNPs (11 primary CD-related and 11 primary RA-related SNPs) were genotyped or imputed in 1368 Dutch RA cases, 795 Dutch CD cases and 1683 controls. Since the distribution of males and females was substantially different in our cases and control cohort (frequency female in both case cohorts 60%, whereas only 20% of controls were female, see Supplementary Material, Table S2), we first compared the allele frequency of all the SNPs in males and females in our control cohort. None of the SNPs showed a significant difference (Supplementary Material, Table S4), which we took as evidence for the comparability of the samples.

Replication of RA loci in Dutch RA cohort

A replication of known RA variants in a Dutch RA cohort has not been performed previously. From a total of 11 RA loci, only the well-established risk variant in PTPN22 (rs2476601) showed a clear association with RA in our cohort ($P = 3.43 \times 10^{-12}$; OR = 1.70 (95% confidence interval (CI): 1.46–1.98)). Three other RA SNPs (rs3890745, rs4810485 and rs3218253 located in the *MMEL1/TNFRSF14, CD40* and the *IL2RB* gene regions, respectively) were nominally

replicated in Dutch RA samples (P < 0.05). The remaining RA loci were not associated in our RA data set (Table 1).

Association of CD loci in Dutch CD cohort

All previously confirmed coeliac variants, except the *LPP* SNP rs1464510, showed association to CD in the Dutch cohort. This is not surprising since we used 508 of the 795 cases, and 833 of the 1683 controls as part of the multi-national cohorts in our previous studies to establish the association of the CD loci[14,20,21].

Cross-disease association study in the Dutch cohorts

Three out of 11 primary CD-related SNPs were associated with RA in the Dutch cohort. The *IL2/IL21* variant rs13151961 showed the strongest association to RA [P=0.0003; OR=0.78 (95% CI: 0.68–0.89)]. This association was reported previously in a subset of our cohort[22]. We now confirmed the association of *IL2/21* variants in an extended group of RA and CD cases. The other two variants, *SH2B3* rs3184504 and *LPP* rs1464510, showed a moderate association with RA [P=0.024; OR = 1.12 (95% CI: 1.02–1.25) and P=0.012; OR=1.14 (95% CI: 1.03–1.26), respectively]. Opposing allelic association for RA and CD was observed for the *LPP* variant: the frequency of *LPP* rs1464510*C allele was increased in RA compared with controls, whereas the rs1464510*A allele was more frequent in CD patients than controls.

From the RA-specific variants, only rs3890745 from the *MMEL1/TNFRSF14* locus showed a trend for association to CD [P=0.04, OR=0.87 (95% CI: 0.77–0.99)]. The rs10499194 variant in the *TNFAIP3* locus, previously reported only in RA samples, also showed moderate association to CD [P=0.018; OR=0.84 (95% CI: 0.73 – 0.97)] (Table 1).

Stratification of RA samples for rheumatoid factor

As the association of several known RA genetic risk variants has been shown to be different in autoantibody positive and negative cases[28], we performed a separate association analysis in rheumatoid factor (RF)-positive and RF-negative cases. We had information on RF status available for a sub- group of Dutch RA cases: 776 cases were RF positive and 204 RF negative. In the RF-positive group the *SH2B3* variant rs3184504 showed stronger association compared to the total RA cohort (P=0.0055, OR=1.19 (95% CI: 1.05– 1.34) (Table 2). We did not perform stratification on anti-CCP auto-antibodies, as the anti-CCP status was only available for a minority of cases.

Table 1. Results of association analysis in Dutch RA and CD cohorts

Gene(s) present in risk locus	Reported association	*rsID*	*CHR*	*Minor allele*	*MAF controls*	RA MAF	*P*-value	OR	95% CI	CD MAF	*P*-value	OR	95% CI
RGS1	*CD*	*rs2816316*	*1*	*C*	*0.19*	*0.18*	*0.797*	*0.98*	*0.86 – 1.12*	*0.14*	*7.37E-05*	*0.72*	*0.61 – 0.84*
REL	*CD*	*rs842647*	*2*	*G*	*0.36*	*0.34*	*0.106*	*0.92*	*0.82 – 1.02*	*0.29*	*1.38E-05*	*0.75*	*0.66 – 0.86*
IL18RAP	*CD*	*rs917997*	*2*	*A*	*0.23*	*0.24*	*0.628*	*1.03*	*0.91 – 1.17*	*0.29*	*1.73E-05*	*1.34*	*1.17 – 1.54*
CCR3	*CD*	*rs6441961*	*3*	*A*	*0.32*	*0.31*	*0.855*	*0.99*	*0.89 – 1.10*	*0.38*	*1.52E-05*	*1.32*	*1.16 – 1.49*
IL12A/SCHIP	*CD*	*rs17810546*	*3*	*G*	*0.12*	*0.11*	*0.613*	*0.96*	*0.82 – 1.13*	*0.17*	*2.73E-07*	*1.55*	*1.31 – 1.84*
IL12A	*CD*	*rs9811792*	*3*	*G*	*0.45*	*0.44*	*0.566*	*0.97*	*0.88 – 1.08*	*0.50*	*0.002*	*1.21*	*1.08 – 1.37*
LPP	*CD*	*rs1464510*	*3*	*C*	*0.49*	*0.53*	*0.013*	*1.14*	*1.03 – 1.26*	*0.47*	*0.102*	*0.91*	*0.80 – 1.02*
IL2-21	*CD*	*rs13151961*	*4*	*G*	*0.18*	*0.15*	*0.0003*	*0.78*	*0.68 – 0.89*	*0.13*	*5.08E-07*	*0.65*	*0.54 – 0.77*
TAGAP	*CD*	*rs1738074*	*6*	*A*	*0.40*	*0.39*	*0.579*	*0.97*	*0.88 – 1.08*	*0.44*	*0.004*	*1.19*	*1.06 – 1.34*
SH2B3	*CD*	*rs3184504*	*12*	*A*	*0.46*	*0.49*	*0.024*	*1.12*	*1.02 – 1.25*	*0.52*	*1.15E-05*	*1.31*	*1.16 – 1.47*
OLIG3-TNFAIP3	*CD/RA*	*rs2327832*	*6*	*G*	*0.21*	*0.22*	*0.230*	*1.08*	*0.95 – 1.22*	*0.23*	*0.0611*	*1.15*	*0.99 – 1.32*
OLIG3-TNFAIP3	*RA*	*rs10499194*	*6*	*T*	*0.24*	*0.24*	*0.987*	*1.00*	*0.89 – 1.13*	*0.21*	*0.018*	*0.84*	*0.73 – 0.97*
MMEL1/TNFRSF14	*RA*	*rs3890745*	*1*	*G*	*0.32*	*0.29*	*0.014*	*0.87*	*0.78 – 0.97*	*0.29*	*0.041*	*0.87*	*0.77 – 0.99*
PTPN22	*RA*	*rs2476601*	*1*	*A*	*0.10*	*0.16*	*3.43E-12*	*1.70*	*1.46 – 1.98*	*0.10*	*0.604*	*0.95*	*0.78 – 1.16*
STAT4	*RA*	*rs7574865*	*2*	*A*	*0.24*	*0.24*	*0.930*	*1.01*	*0.89 – 1.13*	*0.23*	*0.759*	*0.98*	*0.85 – 1.13*
CDK6	*RA*	*rs42041a*	*7*	*G*	*0.25*	*0.26*	*0.462*	*1.06*	*0.91 – 1.22*	*0.26*	*0.704*	*1.03*	*0.88 – 1.22*
CCL21	*RA*	*rs2812378*	*9*	*G*	*0.35*	*0.37*	*0.167*	*1.08*	*0.97 – 1.20*	*0.36*	*0.453*	*1.05*	*0.93 – 1.19*
TRAF1/C5	*RA*	*rs3761847*	*9*	*G*	*0.44*	*0.45*	*0.829*	*1.01*	*0.91 – 1.12*	*0.44*	*0.860*	*0.99*	*0.88 – 1.12*
PFKFB3/PRKCQ	*RA*	*rs4750316*	*10*	*C*	*0.19*	*0.21*	*0.094*	*1.12*	*0.98 – 1.27*	*0.17*	*0.080*	*0.87*	*0.74 – 1.02*
KIF5A	*RA*	*rs1678542*	*12*	*G*	*0.33*	*0.34*	*0.331*	*1.06*	*0.95 – 1.18*	*0.32*	*0.670*	*0.97*	*0.86 – 1.11*
CD40	*RA*	*rs4810485*	*20*	*T*	*0.24*	*0.21*	*0.015*	*0.86*	*0.76 – 0.97*	*0.25*	*0.347*	*1.07*	*0.93 – 1.23*
IL2RB	*RA*	*rs3218253*	*22*	*A*	*0.27*	*0.30*	*0.030*	*1.13*	*1.01 – 1.27*	*0.29*	*0.270*	*1.08*	*0.94 – 1.23*

MAF, minor allele frequency; CHR, chromosome; OR odds ratio. *rs42041 was genotyped in a subgroup of controls (n = 833).

Table 2. Association of RA and CD SNPs in rheumatoid factor-positive subgroup of RA patients (n=775), compared to controls (n=1683)

Gene(s) present in risk locus	CD/RA SNP	rsID	CHR	Minor allele	MAF_cont	MAF RF pos	*P*-value	*OR*	*95% CI*
RGS1	*CD*	*rs2816316*	1	C	0.19	0.18	*0.640*	*0.96*	*0.82 – 1.13*
REL	*CD*	*rs842647*	2	G	0.36	0.33	*0.070*	*0.89*	*0.78 – 1.01*
IL18RAP	*CD*	*rs917997*	2	A	0.23	0.25	*0.234*	*1.09*	*0.95 – 1.26*
CCR3	*CD*	*rs6441961*	3	A	0.32	0.32	*0.824*	*1.02*	*0.89 – 1.16*
IL12A/SCHIP	*CD*	*rs17810546*	3	G	0.12	0.11	*0.693*	*0.96*	*0.80 – 1.17*
IL12A	*CD*	*rs9811792*	3	G	0.45	0.42	*0.066*	*0.89*	*0.79 – 1.01*
LPP	*CD*	*rs1464510*	3	C	0.49	0.52	*0.161*	*1.09*	*0.97 – 1.23*
IL2-21	*CD*	*rs13151961*	4	G	0.18	0.14	*0.0004*	*0.74*	*0.62 – 0.87*
TAGAP	*CD*	*rs1738074*	6	A	0.40	0.38	*0.104*	*0.90*	*0.80 – 1.02*
SH2B3	*CD*	*rs3184504*	12	A	0.46	0.50	*0.0055*	*1.19*	*1.05 – 1.34*
OLIG3-TNFAIP3	*CD/RA*	*rs2327832*	6	G	0.21	0.21	*0.990*	*1.00*	*0.86 – 1.16*
OLIG3-TNFAIP3	*CD/RA*	*rs10499194*	6	T	0.24	0.25	*0.602*	*1.04*	*0.90 – 1.20*
MMEL1-TNFRSF14	*RA*	*rs3890745*	1	G	0.32	0.29	*0.016*	*0.85*	*0.75 – 0.97*
PTPN22	*RA*	*rs2476601*	1	A	0.10	0.17	1.45^{-11}	*1.81*	*1.52 – 2.15*
STAT4	*RA*	*rs7574865*	2	A	0.24	0.24	*0.982*	*1.00*	*0.87 – 1.15*
CDK6	*RA*	*rs42041*[a]	7	G	0.25	0.25	*0.818*	*0.98*	*0.83 – 1.16*
CCL21	*RA*	*rs2812378*	9	G	0.35	0.37	*0.150*	*1.10*	*0.97 – 1.24*
TRAF1/C5	*RA*	*rs3761847*	9	G	0.44	0.45	*0.712*	*1.02*	*0.91 – 1.16*
PFKFB3/PRKCQ	*RA*	*rs4750316*	10	C	0.19	0.22	*0.024*	*1.19*	*1.02 – 1.38*
KIF5Aa	*RA*	*rs1678542*	12	G	0.33	0.34	*0.322*	*1.07*	*0.94 – 1.21*
CD40	*RA*	*rs4810485*	20	T	0.24	0.23	*0.413*	*0.94*	*0.82 – 1.09*
IL2RB	*RA*	*rs3218253*	22	A	0.27	0.29	*0.245*	*1.08*	*0.95 – 1.24*

CHR, chromosome; RF, rheumatoid factor; MAF, minor allele frequency; OR, odds ratio. [a]rs42041 was genotyped in a subgroup of controls (n = 833).

Meta-analysis

To increase the power of the study, we combined our data with those of a GWAS in RA and CD patients from the UK into a meta-analysis. The UK cohorts included 1860 cases and 2938 controls from the WTCCC study on RA, and 767 CD cases and 1422 controls from a UK GWAS in CD[13,21]. In the meta-analysis, four risk loci showed association to both RA and CD with P < 0.01, including the genes *IL2/IL21*, *LPP*, *TNFAIP3* and *SH2B3* (Table 3). Association with the *LPP* locus in CD and RA was observed for opposing alleles. A trend for association with SNPs in *MMEL1-TNFRSF14* and *PRKCQ* was observed for both diseases (P < 0.05) (Table 3). An overview of the common and separate associations of the tested genes with RA and CD is shown in Figure 1.

DISCUSSION

In this study, we (i) performed the replication of known RA associated loci in our Dutch cohort of RA cases and controls; (ii) performed the cross-study of RA and CD associated variants in Dutch RA and CD cohorts and (iii) combined our results in a meta-analysis with the UK GWAS in RA and CD[13,21]. We were able to replicate 4 out of 11 RA loci in our Dutch RA cohort, and identified 6 loci which showed shared association to CD and RA in the meta-analysis. Strikingly, most of the previously established RA SNPs were not replicated. There are several reasons for this observation. First, the endophenotyping difference might be the major reason for non-replication: several of the RA loci were established mainly in subgroups of ACPA-positive RA patients. Differential association in auto-antibody positive and negative subgroups of RA has been reported previously[28]. We could not stratify for ACPA in our cohort due to lack of ACPA status for most patients. However, when stratified for RF, we were able to observe a stronger effect for the variant in the *SH2B3* gene. Secondly, population heterogeneity might be another explanation for lack of association with the initial reported variants. To control for the heterogeneity of the association test, we included the Breslow–Day test in our analysis. Two of the SNPs (rs4750316 and rs1678542, the *PFKFB3/PRKCQ* and *KIF5A* locus, respectively) showed significant results for this test, indicating the presence of heterogeneity between the two populations. Further replication in other populations is essential for establishing the true risk effect of these genes. Since we only tested a single marker (the most associated one for each locus), we may have

Table 3. Meta-analysis of SNPs in Dutch and UK RA and CD populations

Locus	CD/RA SNP	rsID	Ref allele	RA pCMH	OR	95% CI	CD pCMH	OR	95% CI
RGS1	CD	rs2816316	C	0.6349	0.98	0.90- 1.06	1.76E2-07	0.73	0.65 – 0.82
REL	CD	rs842647	G	0.1470	0.95	0.89 – 1.02	2.87E2-06	0.80	0.73 – 0.88
IL18RAP	CD	rs917997	A	0.1071	1.06	0.99 – 1.15	8.62E2-09	1.34	1.21 – 1.48
CCR3	CD	rs6441961	A	0.2672	1.04	0.97 – 1.11	2.46E2-07	1.27	1.16 – 1.39
IL12A/SCHIP	CD	rs17810546	G	0.4926	0.97	0.88 – 1.07	1.45E2-09	1.47	1.30 – 1.67
IL12A	CD	rs9811792	G	0.9945	1.00	0.94 – 1.07	1.16E2-05	1.21	1.11 – 1.32
LPP	CD	rs1464510	C	0.0054	1.10	1.03 – 1.17	0.00010	0.84	0.77 – 0.92
IL2-21	CD	rs13151961	G	1E-05	0.83	0.76 – 0.90	6.61E2-12	0.65	0.58 – 0.74
TAGAP	CD	rs1738074	A	0.0839	0.94	0.88 – 1.01	1.81E2-05	1.21	1.11 – 1.32
SH2B3	CD	rs3184504	A	0.0011	1.11	1.04 – 1.19	4.42E2-08	1.27	1.17 – 1.39
OLIG3-TNFAIP3	CD/RA	rs2327832	G	4E-05	1.17	1.09 – 1.27	0.00069	1.19	1.08 – 1.32
OLIG3-TNFAIP3	CD/RA	rs10499194	T	0.0419	0.93	0.86 – 1.00	0.0419	0.93	0.86 – 1.00
MMEL1-TNFRSF14	RA	rs3890745	G	5E-07	0.84	0.78 – 0.90	0.0275	0.90	0.82 – 0.99
PTPN22	RA	rs2476601	A	2E-27	1.67	1.52 – 1.84	0.2623	1.08	0.94 – 1.25
STAT4	RA	rs7574865	A	0.1041	1.06	0.99 – 1.15	0.3319	1.05	0.95 – 1.16
CDK6	RA	rs42041a	G	0.008	1.11	1.03 – 1.20	n/a	n/a	n/a
CCL21	RA	rs2812378	G	0.0007	1.12	1.05 – 1.20	0.3189	1.05	0.96 – 1.15
TRAF1/C5	RA	rs3761847	G	0.9508	1.00	0.94 – 1.06	0.6610	0.98	0.90 – 1.07
PFKFB3/PRKCQ	RA	rs4750316	C	0.0390	0.92	0.85 – 1.00	0.0135	0.87	0.77 – 0.97
KIF5Aa	RA	rs1678542	G	0.0076	0.91	0.85 – 0.98	0.8002	0.99	0.90 – 1.08
CD40	RA	rs4810485	T	0.0033	0.89	0.83 – 0.96	0.2007	1.07	0.97 – 1.18
IL2RB	RA	rs3218253	A	2E-05	1.17	1.09 – 1.26	0.0568	1.10	1.00 – 1.21

pCMH, P-value Mantel–Haenszel chi-square.
[a]The meta-analysis for rs42041 was not done due to poor imputation of this SNP in the UK and Dutch GWAS data sets.

missed the association due to low/other linkage disequilibrium (LD) between the established 'tagging' SNP and the true causal variant in the Dutch population. Finally, the relative risk of associated regions in the initial studies was rather low (OR between 0.75 and 1.32 for most SNPs). Moreover, similar to the single gene association studies, the effect of some of the established loci might be overestimated and our study might not have had enough power to replicate all loci. Depending on the SNP, the power of our study ranged from 0.2 – 1 in the Dutch cohorts and increased in the meta-analysis to a range of 0.34 – 1 (Supplementary Material, Table S5). The power was at 80% or above for 16 out of 22 SNPs for the RA NL-UK combined analysis and above 80% for 8 out of 22 SNPs for the CD NL-UK analysis. All these factors, or combinations of them, might explain the negative results for most of the genes we tested. Further replication studies in various populations and endophenotypes would shed light on the effect and heterogeneity of associated loci in RA.

In the combined Dutch and UK data sets, we observed a convincing shared association of four genes to RA and CD ($P < 0.01$ in both diseases), and suggestive association of two more loci ($P < 0.05$ in both diseases). Two of the shared loci, *IL2/IL21* and *TNFAIP3*, are already known to be involved in both diseases and were confirmed in our study.

Both genes show association with several other immune-related diseases: *IL2/IL21* with inflammatory bowel disease[29–31], type 1 diabetes[22,24] and psoriasis[23], and *TNFAIP3* with systemic lupus erythematosus[26], type 1 diabetes[25] and psoriasis[32]. This points to a general role for these genes in the development of autoimmunity.

The association of *IL2-IL21* locus to several immune-related diseases is especially interesting, although strong LD in this locus makes it difficult to locate the true associated gene. Both *IL2* and *IL21* are important in T- regulatory and Th17 cells, respectively. Both CD and RA show high levels of Th17 in affected tissues[33,34]. Interestingly, IL21 is produced by Th17 cells and is important for maintaining Th17 cells[35]. High levels of IL21 have been found in the intestine of patients with CD[21], whereas IL21 receptor is overexpressed in synovial tissues of RA patients[36]. Blockading the IL21/IL21R pathway ameliorates disease in a murine model of RA[37].

The association of *SH2B3* with both RA and CD has not been reported previously. The *SH2B3* rs3184504 variant is a non-synonymous SNP R262W located in exon 3 of the gene, and has previously been associated with CD, type 1 diabetes and myocardial infarction[14,24,38]. It encodes the T-cell adapter protein LNK, which regulates T-cell receptor-, growth factor- and cytokine receptor-mediated signalling, and is therefore an attractive candidate gene for shared autoimmune susceptibility[39].

Another shared gene identified in this study, *LPP* (LIM domain containing preferred translocation partner in lipoma), is involved in cell adhesion, cytoskeletal remodelling and maintaining cell shape and motility[40,41]. Chromosomal aberrations including the *LPP* region have been observed in leukaemia, indicating a potential role for this chromosomal region in regulating the immune system[42]. The exact function of *LPP* in autoimmunity has not yet been defined. Interestingly, in our study, LPP shows a differential association for RA and CD. The effect of *LPP* rs1464510*A allele confers protection in RA, and susceptibility in CD. The reason for the opposing allelic associations could be the presence of distinct RA- and CD-causing variants, located on different haplotypes and tagged by the opposite alleles of the same SNP. Another possibility is that the same variant confers truly susceptibility to one disease and protection from another, similar to the *PTPN22* functional variant Arg620Trp, which confers susceptibility to several autoimmune diseases, but protection from Crohn's disease[43,44]. Sequencing of the whole associated block in both diseases is required to define the RA and CD causal variants and explain the exact nature of this observation. Association of opposite alleles to different autoimmunities has been also observed for *IL18RAP* and *TAGAP* variants in CD and type 1 diabetes[45].

Two loci showed moderate association to both diseases- the *PFKFB3/PRKCQ* and *MMEL1-TNFRSF14* ($P < 0.05$). The *PFKFB3/PRKCQ* variant was originally reported in a meta-analysis including RA patients[17]. In addition, the same locus has recently been associated with type 1 diabetes in a meta-analysis[46]. Thus, our findings, although only moderately significant, provide additional support for this locus being a shared autoimmune gene. *PRKCQ* is involved in regulating and controlling T-cell-mediated signalling and is therefore a plausible candidate for autoimmune traits[47]. The *MMEL1/TNFRSF14* locus includes the *TNFRSF14* (HVEM, herpes virus entry mediator) gene, which functions as a co-stimulatory molecule

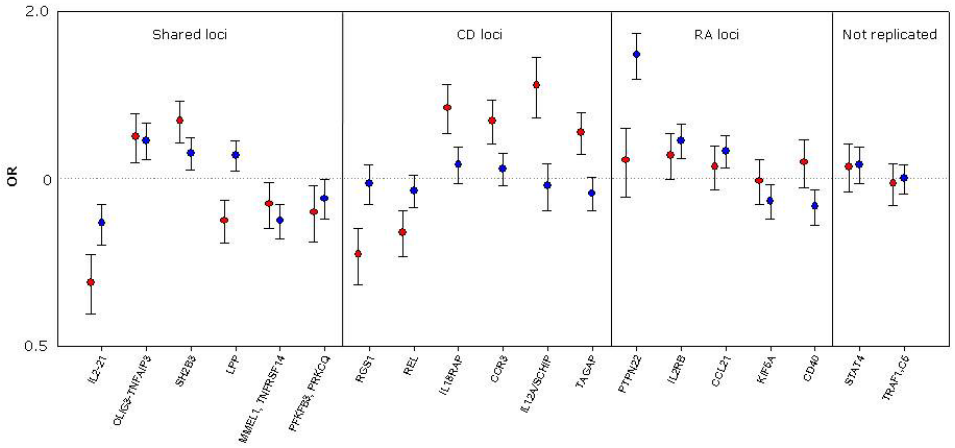


Figure 1. Association of 20 investigated loci to CD and RA. From the *IL12A-SHIP* locus and *OLIG3-TNFAIP3* locus, one of two SNPs (the most associated one, rs17810546 and rs2327832, respectively) are included in the figure. The SNP rs42041 was excluded from the meta-analysis due to poor imputation quality. Blue dots and corresponding line-OR and 95% CI for RA. Red dots and lines-OR and 95% CI for CD. OR was calculated for the minor allele of associated SNP.

during T-cell activation[48], and enhances the bactericidal activities of human monocytes and neutrophils[49]. In antiviral responses, *TNFRSF14* is involved in NF-κB activation.

Overall, from six shared CD-RA loci, five contain genes directly involved in immune function. Association to the *SH2B3, PRKCQ, IL2/21* and *TNFRSF14* genes points to the role of T-cell-mediated signalling, whereas both the *TNFAIP3* and *TNFRSF14* genes are linked to NF-κb signalling, innate immunity and the response to pathogens. The association of innate molecules to RA and CD may now explain the link of mucosal immunity state and infections in predisposition to both diseases[50-52].

Our study extends the knowledge on genes that are shared between RA and CD. Most of the shared genes fit with the current hypotheses on the pathogenesis of these diseases, involving both innate and adaptive immunity. Several of the shared genes also contribute to the susceptibility to additional autoimmune diseases. Creating genetic profiles for autoimmunity may help understand the basic mechanisms of pathogenesis and predict common immune-related risk factors/ phenotypes. Furthermore, it will open up possibilities for the development of new drug targets for a better treatment of autoimmune diseases.

MATERIALS AND METHODS

Study population

Rheumatoid arthritis cohorts. We combined two independent Dutch RA inception cohorts: from Nijmegen (n=960) and Groningen (n=408). Both cohorts have been described elsewhere[22,53,54]. All patients were diagnosed according to the American College of Rheumatology (ACR) criteria for RA[55]. Due to the lack of whole genome data, we ascertained subjects were of Dutch descent based on their surname.

Coeliac disease cohorts. Our study analysed 795 unrelated Dutch individuals with CD. All affected individuals were diagnosed according to the revised ESPGAN criteria[56]. The cohort encompassed individuals that showed a Marsh II or Marsh III lesion in the initial diagnostic small-bowel biopsy specimens upon re-evaluation by one of two experienced pathologists, or presented with dermatitis herpetiformis and were HLA-DQ2 positive.

Control cohort. The control cohort comprised unrelated blood bank donors (n=833) and NELSON controls (n=850). The blood bank control cohort was described earlier[14,20]. Other controls were included from the NELSON project-an ongoing population-based, randomized multi-centre lung cancer screening trial, studying male smokers[57]. These controls were collected from the north and centre of the Netherlands (Groningen, Utrecht and Drenthe, The Netherlands). All the control subjects were heavy smokers or ex-smokers (a minimum of 16 cigarettes/day for 25 years or 11 cigarettes/ day for 30 years), but did not develop airway obstruction or emphysema suggesting chronic obstructive pulmonary disease (COPD) until the end of a 4 year observation period.

The current study was approved by the local ethics committees and all the patients and controls gave their written informed consent. Information on male/female ratio of the cohorts is shown in Supplementary Material, Table S2.

SNP selection and genotyping

We selected SNPs with confirmed association to either RA or CD with $P < 5 \times 10^{-06}$ from existing literature. Information on the original studies and the associated SNPs is given in Supplementary Material, Table S1. In total, we tested 22 SNPs from 20 loci, 11 known for association to CD and 11 primarily associated to RA, with two of the genes known to be associated to diseases.

Genotyping of Dutch RA samples. All 22 SNPs were genotyped in the Dutch RA cohort using TaqMan probes and primers developed by Applied Biosystems, on an ABI 7900HT system (Applied Biosystems, Nieuwerkerk a/d IJssel,

The Netherlands). Genotyping was performed following the manufacturer's specifications. DNA samples were processed in 384-well plates, each plate contained 8 negative controls and 5 duplicated samples. All duplicates showed consistent results for all SNPs.

Genotyping of Dutch CD samples and controls. Of the 22 SNPs, 15 were present on the Illumina HAP550 platform. The genotyping data of CD cases and controls for these 15 SNPs were extracted from an ongoing GWAS in the Dutch population, performed on the Human670-QuadCustom Illumina BeadChips, which contains the HumanHap550 SNP set (manuscript in preparation). The remaining seven SNPs were not present on the genotyping platforms. In these cases, a 4 MB window around the SNP of interest was imputed using Plink v1.05 (http://pngu.mgh.harvard.edu/purcell/plink/) and the phase 2, HapMap CEU samples, release 23a, as the imputation reference panel[58]. We included the imputed genotypes only if the imputation quality was above 0.8 (Supplementary Material, Table S3). The SNP rs42041 showed deviation from Hardy–Weinberg equilibrium (HWE) in the Dutch GWAS imputed data set, so this SNP was genotyped by TaqMan in Dutch CD cases and the blood bank control group (n=833). NELSON controls (n=850) were not genotyped for the rs42041. TaqMan genotyping was performed following the manufacturer's specifications. DNA samples were processed in 384-well plates; each plate contained 8 negative controls and 16 genotyping controls [four duplicates of four different samples obtained from the Centre d'Etude du Polymorphisme Humain (CEPH)].

UK data sets. For RA, we used freely available genotype data for 9 SNPs from RA patients from the GWAS preformed by the Wellcome Trust Case Control Consortium (WTCCC) in 2007[13], comprising 1860 UK RA cases and 2938 UK controls. For the remaining 13 SNPs, we used imputed genotypes, all with imputation quality > 0.8 (https://www.wtccc.org.uk/ccc1/summary_stats.shtml data access 21 August 2008).

For CD, genotyping results for 15 SNPs were extracted from the UK GWAS study[21]. The non-genotyped SNPs were imputed in the UK cases and controls (imputation quality (info) > 0.8, except for the SNP rs42041, which could not be imputed with sufficient quality and was therefore excluded from the meta-analysis (Supplementary Material, Table S3).

The genotyping method and imputation quality in each cohort is presented in Supplementary Material, Table S3A-C. The frequency of missing genotypes was below 5% for all the SNPs in all the genotyped cohorts.

Population substructure analysis

Multidimensional scaling analysis (implemented in Plink) was applied to all the samples for which genome-wide genotype data were available. The values of the first five components were further plotted against each other. We detected outlying samples only for the first three components. Samples exceeding two standard deviations from the mean for the first three components were excluded, ensuring that only the most homogeneous samples were included in our final analysis and that there was no population stratification.

As a part of the ongoing GWAS in celiac disease, we applied a multidimensional scaling (MDS) analysis to assess whether there was a large population substructure between the UK and Dutch cohorts. Both cohorts created tight clusters and, as expected, the two first components could distinguish between the two populations. However, over 40% of the samples shared a common part in the plot. Both cohorts were also merged with the HapMap2 data and the MDS analysis was repeated. Both cohorts mapped perfectly to the CEU population, any outlying samples that mapped outside the cluster were excluded from further analysis.

Statistical analysis

We calculated HWE for all the genotyped SNPs by comparing the expected and observed

genotypes in a 2x3 chi-squared table. None of the markers deviated significantly from HWE in cases or controls (P > 0.01) in any of the populations. Association analysis was performed using chi-squared statistics with two-tailed P-values (implemented in Plink v1.05)[58]. In the meta-analysis, we combined the information on allele counts for the Dutch and UK cohorts separately for CD and RA, in the Mantel-Haenszel chi-squared association test with two clusters. P-values, odds ratios (ORs) and 95% confidence intervals (95% CIs) were calculated using Plink v1.05.

To estimate the heterogeneity of the association tests between the populations, we performed the Breslow-Day test. All except two SNPs described in this study were negative for this test. Rs4750316 and rs1678542 showed significant results for the Breslow-Day test in the RA cohort and we therefore applied the random-effect model for non-heterogeneous studies (the DerSimonia-Laird method implemented in the rmeta package for R; www.r-project.org).

Power calculation

We calculated the power of our sample size to detect significant associations using the Genetic Power Calculator (http://pngu.mgh.harvard.edu/~purcell/gpc/)[59]. The calculations were performed separately for the Dutch cohort and for the meta-analysis with UK collections and are summarized in Supplementary Material, Table S5.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

G.T. was awarded a Ter Meulen Fund travel grant by the Royal Netherlands Academy of Arts and Sciences (KNAW). We thank all the individuals who participated in the study, Jackie Senior for critically reading the manuscript, and Flip Mulder for help with graphic design.

Conflict of Interest statement. None declared.

FUNDING

The study was supported by the Celiac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009), by the Netherlands Organization for Scientific Research (NWO, VICI grant 918.66.620 to CW), the Well- come Trust (WT084743MA). We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant Goooo934 and the Well- come Trust grant 068545/Z/02. M.C. was supported by a per- sonal grant from the Netherlands Organization for Scientific Research (NOW, VENI grant 916.76.020). GT was awarded a Ter Meulen Fund travel grant by the Royal Netherlands Academy of Arts and Sciences (KNAW). We thank all the individuals who participated in the study, Jackie Senior for critically reading the manuscript, and Flip Mulder for help with graphic design.

REFERENCES

1. Oliver, J.E., Worthington, J. and Silman, A.J. (2006) Genetic epidemiology of rheumatoid arthritis. Curr. Opin. Rheumatol., 18, 141 – 146.

2. Oliver, J.E. and Silman, A.J. (2006) Risk factors for the development of rheumatoid arthritis. Scand. J. Rheumatol., 35, 169–174.

3. Somers, E.C., Thomas, S.L., Smeeth, L. and Hall, A.J. (2006) Autoimmune diseases co-occurring within individuals and within families: a systematic review. Epidemiology, 17, 202–217.

4. Zhernakova, A., van Diemen, C.C. and Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet., 10, 43–55.

5. Sollid, L.M. (2002) Coeliac disease: dissecting a complex inflammatory disorder. Nat. Rev. Immunol., 2, 647–655.

6. Molberg, O. and Sollid, L.M. (2006) A gut feeling for joint inflammation—using coeliac disease to understand rheumatoid arthritis. Trends Immunol., 27, 188 – 194.

7. Bourne, J.T., Kumar, P., Huskisson, E.C., Mageed, R., Unsworth, D.J. and Wojtulewski, J.A. (1985) Arthritis and coeliac disease. Ann. Rheum. Dis., 44, 592 – 598.

8. Collin, P., Korpela, M., Hallstrom, O., Viander, M., Keyrilainen, O. and Maki, M. (1992) Rheumatic complaints as a presenting symptom in patients with coeliac disease. Scand. J. Rheumatol., 21, 20 – 23.

9. Neuhausen, S.L., Steele, L., Ryan, S., Mousavi, M., Pinto, M., Osann, K.E., Flodman, P. and Zone, J.J. (2008) Co-occurrence of celiac disease and other autoimmune diseases in celiacs and their first-degree relatives. J. Autoimmun., 31, 160–165.

10. Parke, A.L., Fagan, E.A., Chadwick, V.S. and Hughes, G.R. (1984) Coeliac disease and rheumatoid arthritis. Ann. Rheum. Dis., 43, 378–380.

11. Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Gibbons, L., Plant, D., Wilson, A.G., Marinou, I. et al. (2008) Re-evaluation of putative rheumatoid arthritis susceptibility genes in the post-genome wide association study era and hypothesis of a key pathway underlying susceptibility. Hum. Mol. Genet., 17, 2274–2279.

12. Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Plant, D., Gibbons, L.J., Wilson, A.G., Bax, D.E. et al. (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat. Genet., 40, 1156–1159.

13. Consortium, W.T.C.C. (2007) Genome-wide association

study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447, 661 – 678.

14. Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D. et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet., 40, 395–402.

15. Plenge, R.M., Cotsapas, C., Davies, L., Price, A.L., de Bakker, P.I., Maller, J., Pe'er, I., Burtt, N.P., Blumenstiel, B., DeFelice, M. et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat. Genet., 39, 1477–1482.

16. Plenge, R.M., Seielstad, M., Padyukov, L., Lee, A.T., Remmers, E.F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L.R. et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis—a genome-wide study. N. Engl. J. Med., 357, 1199–1209.

17. Raychaudhuri, S., Remmers, E.F., Lee, A.T., Hackett, R., Guiducci, C., Burtt, N.P., Gianniny, L., Korman, B.D., Padyukov, L., Kurreeman, F.A. et al. (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat. Genet., 40, 1216–1223.

18. Remmers, E.F., Plenge, R.M., Lee, A.T., Graham, R.R., Hom, G., Behrens, T.W., de Bakker, P.I., Le, J.M., Lee, H.S., Batliwalla, F. et al. (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N. Engl. J. Med., 357, 977–986.

19. Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J., Donn, R., Symmons, D., Hider, S., Bruce, I.N. et al. (2007) Rheumatoid arthritis association at 6q23. Nat. Genet., 39, 1431–1433.

20. Trynka, G., Zhernakova, A., Romanos, J., Franke, L., Hunt, K., Turner, G., Platteel, M., Ryan, A.W., de Kovel, C., Barisani, D. et al. (2009) Coeliac disease associated risk variants in TNFAIP3 and REL implicate altered NF-fkappagB signalling. Gut., 58, 1078 – 1083.

21. van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K. et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet., 39, 827 – 829.

22. Zhernakova, A., Alizadeh, B.Z., Bevova, M., van Leeuwen, M.A., Coenen, M.J., Franke, B., Franke, L., Posthumus, M.D., van Heel, D.A., van der Steege, G. et al. (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am. J. Hum. Genet., 81, 1284 – 1288.

23. Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C., Miner, A., Malloy, M.J., Pullinger, C.R. et al. (2008) A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet., 4, e1000041.

24. Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F. et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet., 39, 857–864.

25. Fung, E.Y., Smyth, D.J., Howson, J.M., Cooper, J.D., Walker, N.M., Stevens, H., Wicker, L.S. and Todd, J.A. (2009) Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/ TNFAIP3 as a susceptibility locus. Genes Immun., 10, 188 – 191.

26. Graham, R.R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C.J., Leon, J.M., Burtt, N.P., Guiducci, C., Parkin, M., Gates, C. et al. (2008) Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat. Genet., 40, 1059 – 1061.

27. Musone, S.L., Taylor, K.E., Lu, T.T., Nititham, J., Ferreira, R.C., Ortmann, W., Shifrin, N., Petri, M.A., Kamboh, M.I., Manzi, S. et al. (2008) Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nat. Genet., 40, 1062 – 1064.

28. van der Helm-van Mil, A.H. and Huizinga, T.W. (2008) Advances in the genetics of rheumatoid arthritis point to subclassification into distinct disease subsets. Arthritis Res. Ther., 10, 205.

29. Festen, E.A., Goyette, P., Scott, R., Annese, V., Zhernakova, A., Lian, J., Lefebvre, C., Brant, S.R., Cho, J.H., Silverberg, M.S. et al. (2009) Genetic variants in the region harbouring IL2/IL21 associated with ulcerative colitis. Gut, 58, 799–804.

30. Glas, J., Stallhofer, J., Ripke, S., Wetzke, M., Pfennig, S., Klein, W., Epplen, J.T., Griga, T., Schiemann, U., Lacher, M. et al. (2009) Novel genetic risk markers for ulcerative colitis in the IL2/IL21 region are in epistasis with IL23R and suggest a common genetic background for ulcerative colitis and celiac disease. Am. J. Gastroenterol., 104,1737 – 1744.

31. Marquez, A., Orozco, G., Martinez, A., Palomino-Morales, R., Fernandez-Arquero, M., Mendoza, J.L., Taxonera, C., Diaz-Rubio, M., Gomez-Garcia, M., Nieto, A. et al. (2009) Novel association of the Interleukin 2-Interleukin 21 Region with inflammatory bowel disease. Am. J. Gastroenterol. [Epub ahead of print].

32. Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E., Li, Y., Tejasvi, T., Feng, B.J. et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet., 41, 199–204.

33. Castellanos-Rubio, A., Santin, I., Irastorza, I., Castano, L., Carlos Vitoria, J. and Ramon Bilbao, J. (2009) TH17 (and TH1) signatures of intestinal biopsies of CD patients in response to gliadin. Autoimmunity, 42, 69–73.

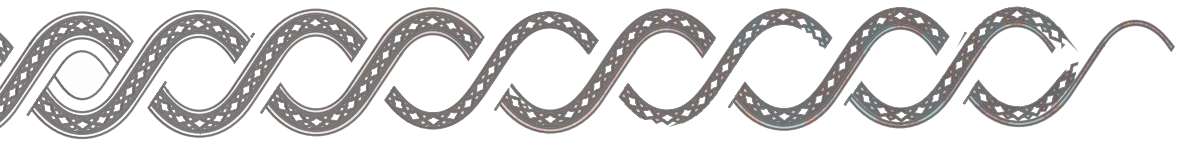
34. Pernis, A.B. (2009) Th17 cells in rheumatoid arthritis and systemic lupus erythematosus. J. Intern. Med., 265, 644–652.

35. Deenick, E.K. and Tangye, S.G. (2007) Autoimmunity: IL-21: a new player in Th17-cell differentiation. Immunol. Cell. Biol., 85, 503–505.

36. Jungel, A., Distler, J.H., Kurowska-Stolarska, M., Seemayer, C.A., Seibl, R., Forster, A., Michel, B.A., Gay, R.E., Emmrich, F., Gay, S. et al. (2004) Expression of interleukin-21 receptor, but not interleukin-21, in synovial fibroblasts and synovial macrophages of patients with rheumatoid arthritis. Arthritis Rheum., 50, 1468–1476.

37. Young, D.A., Hegen, M., Ma, H.L., Whitters, M.J., Albert, L.M., Lowe, L., Senices, M., Wu, P.W., Sibley, B., Leathurby, Y. et al. (2007) Blockade of the interleukin-21/ interleukin-21 receptor pathway ameliorates disease in animal models of rheumatoid arthritis. Arthritis Rheum.,

56, 1152–1163.
38. Gudbjartsson, D.F., Bjornsdottir, U.S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G.M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H. et al. (2009) Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat. Genet., 41, 342 – 347.
39. Fitau, J., Boulday, G., Coulon, F., Quillard, T. and Charreau, B. (2006) The adaptor molecule Lnk negatively regulates tumor necrosis factor-alpha-dependent VCAM-1 expression in endothelial cells through inhibition of the ERK1 and -2 pathways. J. Biol. Chem., 281,20148 – 20159.
40. Jin, L., Kern, M.J., Otey, C.A., Wamhoff, B.R. and Somlyo, A.V. (2007) Angiotensin II, focal adhesion kinase, and PRX1 enhance smooth muscle expression of lipoma preferred partner and its newly identified binding partner palladin to promote cell migration. Circ. Res., 100, 817 – 825.
41. Petit, M.M., Meulemans, S.M. and Van de Ven, W.J. (2003) The focal adhesion and nuclear targeting capacity of the LIM-containing lipoma-preferred partner (LPP) protein. J. Biol. Chem., 278, 2157 – 2168.
42. Daheron, L., Veinstein, A., Brizard, F., Drabkin, H., Lacotte, L., Guilhot, F., Larsen, C.J., Brizard, A. and Roche, J. (2001) Human LPP gene is fused to MLL in a secondary acute leukemia with a t(3;11) (q28;q23). Genes Chromosomes Cancer, 31, 382 – 389.
43. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet., 40, 955 – 962.
44. Vang, T., Miletic, A.V., Bottini, N. and Mustelin, T. (2007) Protein tyrosine phosphatase PTPN22 in human autoimmunity. Autoimmunity, 40, 453 – 461.
45. Smyth, D.J., Plagnol, V., Walker, N.M., Cooper, J.D., Downes, K., Yang, J.H., Howson, J.M., Stevens, H., McManus, R., Wijmenga, C. et al. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med., 359, 2767–2777.
46. Cooper, J.D., Smyth, D.J., Smiles, A.M., Plagnol, V., Walker, N.M., Allen, J.E., Downes, K., Barrett, J.C., Healy, B.C., Mychaleckyj, J.C. et al. (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat. Genet., 40, 1399–1401.
47. Hayashi, K. and Altman, A. (2007) Protein kinase C theta (PKCtheta): a key player in T cell life and death. Pharmacol. Res., 55, 537–544.
48. Gonzalez, L.C., Loyet, K.M., Calemine-Fenaux, J., Chauhan, V., Wranik, B., Ouyang, W. and Eaton, D.L. (2005) A coreceptor interaction between the CD28 and TNF receptor family members B and T lymphocyte attenuator and herpesvirus entry mediator. Proc. Natl Acad. Sci. USA, 102, 1116–1121.
49. Heo, S.K., Ju, S.A., Lee, S.C., Park, S.M., Choe, S.Y., Kwon, B., Kwon, B.S. and Kim, B.S. (2006) LIGHT enhances the bactericidal activity of human monocytes and neutrophils via HVEM. J. Leukoc. Biol., 79, 330 – 338.
50. Rashid, T. and Ebringer, A. (2008) Rheumatoid arthritis in smokers could be linked to Proteus urinary tract infections. Med. Hypotheses, 70, 975 – 980.
51. Stene, L.C., Honeyman, M.C., Hoffenberg, E.J., Haas, J.E., Sokol, R.J., Emery, L., Taki, I., Norris, J.M., Erlich, H.A., Eisenbarth, G.S. et al. (2006) Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am. J. Gastroenterol., 101, 2333–2340.
52. Vaahtovuo, J., Munukka, E., Korkeamaki, M., Luukkainen, R. and Toivanen, P. (2008) Fecal microbiota in early rheumatoid arthritis. J. Rheumatol., 35, 1500–1505.
53. Toonen, E.J., Coenen, M.J., Kievit, W., Fransen, J., Eijsbouts, A.M., Scheffer, H., Radstake, T.R., Creemers, M.C., de Rooij, D.J., van Riel, P.L. et al. (2008) The tumour necrosis factor receptor superfamily member 1b 676T.G polymorphism in relation to response to infliximab and adalimumab treatment and disease severity in rheumatoid arthritis. Ann. Rheum. Dis., 67, 1174–1177.
54. Welsing, P.M. and van Riel, P.L. (2004) The Nijmegen inception cohort of early rheumatoid arthritis. J. Rheumatol. Suppl., 69, 14–21.
55. Arnett, F.C., Edworthy, S.M., Bloch, D.A., McShane, D.J., Fries, J.F., Cooper, N.S., Healey, L.A., Kaplan, S.R., Liang, M.H., Luthra, H.S. et al. (1988) The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum., 31, 315–324.
56. United European Gastroenterology. (2001) When is a coeliac a coeliac? Report of a working group of the United European Gastroenterology Week in Amsterdam, 2001. Eur. J. Gastroenterol. Hepatol., 13, 1123– 1128.
57. van Iersel, C.A., de Koning, H.J., Draisma, G., Mali, W.P., Scholten, E.T., Nackaerts, K., Prokop, M., Habbema, J.D., Oudkerk, M. and van Klaveren, R.J. (2007) Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int. J. Cancer, 120, 868 – 874.
58. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet., 81, 559 – 575.
59. Purcell, S., Cherny, S.S. and Sham, P.C. (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics, 19, 149 – 150.

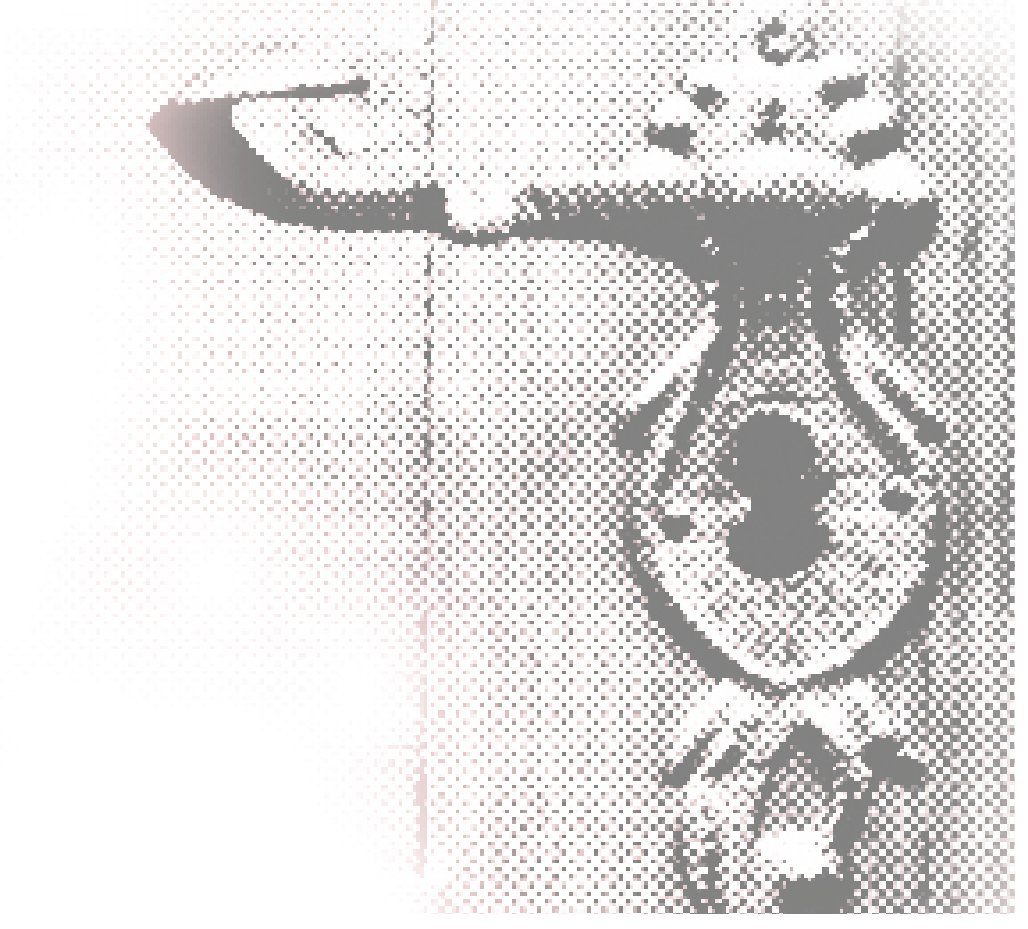


Multiple common variants for celiac disease influencing immune gene expression

Nat Genet. 2010 Apr;42(4):295-302

Gosia Trynka*, Patrick C A Dubois*, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham A R Heap, Róza Ádány, Arpo Aromaa, Maria Teresa Bardella, Leonard H van den Berg, Nicholas A Bockett, Emilio G de la Concha, Bárbara Dema, Rudolf S N Fehrmann, Miguel Fernández-Arquero, Szilvia Fiatal, Elvira Grandone, Peter M Green, Harry J M Groen, Rhian Gwilliam, Roderick H J Houwen, Sarah E Hunt, Katri Kaukinen, Dermot Kelleher, Ilma Korponay-Szabo, Kalle Kurppa, Padraic MacMathuna, Markku Mäki, Maria Cristina Mazzilli, Owen T McCann, M Luisa Mearin, Charles A Mein, Muddassar M Mirza, Vanisha Mistry, Barbara Mora, Katherine I Morley, Chris J Mulder, Joseph A Murray, Concepción Núñez, Elvira Oosterom, Roel A Ophoff, Isabel Polanco, Leena Peltonen, Mathieu Platteel, Anna Rybak, Veikko Salomaa, Joachim J Schweizer, Maria Pia Sperandeo, Greetje J Tack, Graham Turner, Jan H Veldink, Wieke H M Verbeek, Rinse K Weersma, Victorien M Wolters, Elena Urcelay, Bozena Cukrowska, Luigi Greco, Susan L Neuhausen, Ross McManus, Donatella Barisani, Panos Deloukas, Jeffrey C Barrett, Paivi Saavalainen, Cisca Wijmenga & David A van Heel

*equal contribution



ABSTRACT

We performed a second-generation genome-wide association study of 4,533 individuals with celiac disease (cases) and 10,750 control subjects. We genotyped 113 selected SNPs with $P_{GWAS} < 10^{-4}$ and 18 SNPs from 14 known loci in a further 4,918 cases and 5,684 controls. Variants from 423 new regions reached genome-wide significance ($P_{combined} < 5 \times 10^{-8}$); most contain genes with immune functions (*BACH2*, *CCR4*, *CD80*, *CIITA-SOCS1-CLEC16A*, *ICOSLG* and *ZMIZ1*), with *ETS1*, *RUNX3*, *THEMIS* and *TNFRSF14* having key roles in thymic T-cell selection. There was evidence to suggest associations for a further 13 regions. In an expression quantitative trait meta-analysis of 42,469 whole blood samples, 20 of 38 (52.6%) tested loci had celiac risk variants correlated ($P < 0.0028$, FDR 5%) with *cis* gene expression.

Celiac disease is a common heritable chronic inflammatory condition of the small intestine induced by dietary wheat, rye and barley, as well as other unidentified environmental factors, in susceptible individuals. Specific *HLA-DQA1* and *HLA-DQB1* risk alleles are necessary, but not sufficient, for disease development[1,2]. The well-defined role of *HLA-DQ* heterodimers encoded by these alleles is to present cereal peptides to $CD4^+$ T cells, activating an inflammatory immune response in the intestine. A single genome-wide association study (GWAS) has been performed in celiac disease, which identified the *IL2-IL21* risk locus[1].

Subsequent studies probing the GWAS information in greater depth have identified a further 12 risk regions. Most of these regions contain a candidate gene that functions in the immune system, although only in the case of *HLA-DQA1* and *HLA-DQB1* have the causal variants been established[3–5]. Many of the known celiac disease–associated loci overlap with those of other immune-related diseases[6]. To identify additional risk variants, particularly those with smaller effect sizes, we performed a second-generation GWAS using more than six times as many samples as the previous GWAS and a denser genome-wide SNP set. We followed up promising findings in a large collection of independent samples.

RESULTS

Overview of study design

The GWAS included five European celiac disease case and control sample collections, including the celiac disease dataset reported previously[1]. We performed stringent data quality control (see Online Methods), including calling genotypes using a custom algorithm on both large sample sets and, where possible, cases and controls together (see Online Methods). We tested 292,387 non-HLA SNPs from the Illumina Hap300 marker set for association in 4,533 individuals with celiac disease and 10,750 control subjects of European descent (Table 1). A further 231,362 additional non-HLA markers from the Illumina Hap550 marker set were tested for association in a subset of 3,796 individuals with celiac disease and 8,154 controls. All markers were from autosomes or the X chromosome. Genotype call rates were >99.9% in both datasets. The overdispersion factor of association test statistics, $\lambda_{GC} = 1.12$, was similar to that observed in other GWASs of this sample size[7,8]. Findings were not substantially altered by imputation of missing genotypes for 737 cases with celiac disease genotyped on the Hap300 BeadChip and corresponding controls (Table 1, collection 1). Here we present results for directly genotyped SNPs, as around half the additional Hap550 markers cannot be accurately imputed from Hap300 data[9] (including the new *ETS1* locus reported in this study). Results for the top 1,000 markers are available in Supplementary Data 1; however, because of concerns regarding the detection of individuals' identities[10], results for all markers are available only on request to the corresponding author.

For follow-up, we first inspected genotype clouds for the 417 non-HLA SNPs that met $P_{GWAS} < 10^{-4}$, being aware that top GWAS signals might be enriched for genotyping artifact, and excluded 22 SNPs from further analysis using a low threshold for possible bias. We selected SNPs from 113 loci for replication. Markers that passed design and genotyping quality control included (i) 18 SNPs from all 14 previously identified celiac disease risk loci (including a tag SNP for the major celiac disease–associated *HLA-DQ2.5cis* haplotype[1]); (ii) 13 SNPs from all 7 newly discovered regions with $P_{GWAS} < 5 \times 10^{-7}$; (iii) 86 SNPs from 59 of 68 newly discovered regions with $5 \times 10^{-7} < P < 5 \times 10^{-5}$ in stage 1; and (iv) 14 SNPs GWAS from 14 of 30 newly discovered regions with $5 \times 10^{-5} < P_{GWAS} < 10^{-4}$ in stage 1 (for this last category, we mostly chose regions with immune system genes). Two SNPs were selected per region for regions with stronger association, regions with possible multiple independent associations and/or regions containing genes of obvious biological interest. We successfully genotyped 131 SNPs in 7 independent follow-up cohorts comprising 4,918 individuals with

celiac disease and 5,684 control subjects of European descent (Table 1). Genotype call rates were >99.9% in each collection. Primary association analyses of the combined GWAS and follow-up data were performed with a two-sided 2 × 2 × 12 Cochran-Mantel-Haenszel test. Finally, we examined associated risk loci for cis expression-genotype correlations; a summary of subjects used for expression quantitative trait locus (eQTL) analyses is reported in Supplementary Table 1.

Celiac disease risk variants

The HLA locus and all 13 other previously reported celiac disease risk loci showed evidence for association at a genome-wide significance threshold ($P_{combined} < 5 \times 10^{-8}$; Table 2 and Supplementary Fig. 1). We note that some loci were previously reported using less stringent criteria (for example, the $P < 5 \times 10^{-7}$ recommended by the 2007 WTCCC study[11]); however, in the current, much larger sample set, all known loci meet recently proposed $P < 5 \times 10^{-8}$ thresholds[12,13]. We identified 13 new risk regions with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$) of association, including regions containing the *BACH2*, *CCR4*, *CD80*, *CIITA-SOCS1-CLEC16A*, *ETS1*, *ICOSLG*, *RUNX3*, *THEMIS*, *TNFRSF14* and *ZMIZ1* genes, which have obvious immunological functions (Table 2 and Supplementary Fig. 1). A further 13 regions met 'suggestive' criteria for association ($10^{-6} > P_{combined} > 5 \times 10^{-8}$ and/or $P_{GWAS} < 10^{-4}$ and $P_{follow-up} < 0.01$; Table 2 and Supplementary Fig. 1).

Table 1 Sample collections and genotyping platforms

		Celiac disease cases			Controls		
Collection	Country	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]
Stage 1: Genome wide association							
1[d,e]	UK	778	737	Illumina Hap300v1-1	2,596	2,596	Illumina Hap550-2v3
2[d,f]	UK	1,922	1,849	Illumina 670-QuadCustom_v1	5,069	4,936	Illumina 1.2M-DuoCustom_v1
3[d]	Finland	674	647	Illumina 670-QuadCustom_v1	1,839	1,829	Illumina 610-Quad
4[g]	Netherlands	876	803	Illumina 670-QuadCustom_v1	960	846	Illumina 670-QuadCustom_v1
5[d]	Italy	541	497	Illumina 670-QuadCustom_v1	580	543	Illumina 670-QuadCustom_v1
Analysis of Hap300 markers[c]			4,533			10,750	
Analysis of additional Hap550 markers[c]			3,796			8,154	
Stage 2: Follow-up							
6	USA	987	973	Illumina GoldenGate	615	555	Illumina GoldenGate
7	Hungary	979	965	Illumina GoldenGate	1,126	1,067	Illumina GoldenGate
8[h]	Ireland	653	597	Illumina GoldenGate	1,499	1,456	Illumina GoldenGate
9	Poland	599	564	Illumina GoldenGate	745	716	Illumina GoldenGate
10	Spain	558	550	Illumina GoldenGate	465	433	Illumina GoldenGate
11[d]	Italy	1,056	1,010	Illumina GoldenGate	864	804	Illumina GoldenGate
12[d]	Finland	270	259	Illumina GoldenGate	653[j]	653	Illumina GoldenGate[j]
Subtotal			4,918			5,684	
Analysis of Hap300 markers, and follow-up (91 SNPs)[c]			9,451			16,434	
Analysis of additional Hap550 markers, and follow-up (40 SNPs)[c]			8,714			13,838	

[a]Sample numbers attempted for genotyping, before any quality control (QC) steps were applied. [b]Sample numbers after all quality control (QC) steps (used in the association analysis). [c]All platforms contain a common set of Hap300 markers; the Hap550, 610-Quad, 670-Quad and 1.2M contain a common set of Hap550 markers. [d]As an additional quality control step, we performed case-case and control-control comparisons for collection 1 versus 2, and collection 3 versus 12, for the 40 SNPs in table 2 and observed no markers with P < 0.01. We did observe (as expected) differences for collection 5 versus 11, from northern and southern Italy, respectively. [e]All 737 post-QC cases reported in a previous GWAS[1]. [f]690 of the post-QC cases and 1,150 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [g]498 of the post- QC cases and 767 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [h]352 of the post-QC cases and 921 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [i]Some of these data were generated elsewhere, and some prior quality control steps (information not available) had been applied. [j]Finnish stage 2 controls were individuals within the Finrisk collection for whom Illumina 610-Quad genotype data became available after the completion of stage 1.

Table 2 Genomic regions with the strongest association signals for celiac disease

Chr	Position (bp)	SNP	LD block[ab] (Mb)	Minor allele	Minor allele freq[c]	P_{GWAS} 4,533 cases, 10,750 controls	$P_{follow-up}$ 4918 cases, 5684 controls	$P_{combined}$ 9451 cases, 16434 controls	Odds ratio[c] [95% CI]	Multiple independent association signals[d]	Ref	RefSeq Genes in LD block	Genes of interest and GRAIL annotation[e]
Previously reported risk variants													
1	190803436	rs28163316	190.73-190.81	C	0.160	1.45×10^{-12}	1.56×10^{-6}	2.20×10^{-17}	0.80 [0.76-0.84]		22	1	*RGS1*
2	61040333	rs13003464	60.78-61.74	G	0.401	4.92×10^{-8}	1.57×10^{-6}	3.71×10^{-13}	1.15 [1.11-1.20]	yes	32	8	*REL*, *AHSA2*
2	102437000	rs917997	102.22-102.57	A	0.236	5.97×10^{-15}	7.83×10^{-4}	1.11×10^{-15}	1.19 [1.14-1.25]		22	5	*IL18RAP*, *IL18R1*, *IL1RL1*, *IL1RL2*
2	181704290	rs13010713	181.50-181.97	G	0.448	2.02×10^{-8}	3.21×10^{-4}	4.74×10^{-11}	1.13 [1.09-1.18]		33	1	*ITGA4*, *UBE2E3*
2	204510823	rs4675374	204.40-204.52	A	0.223	8.80×10^{-8}	4.94×10^{-3}	5.79×10^{-9}	1.14 [1.09-1.19]		17	2	*CTLA4*, *ICOS*, *CD28*
3	46210205	rs13098911	45.90-46.57	A	0.097	2.53×10^{-11}	1.96×10^{-7}	3.26×10^{-17}	1.30 [1.23-1.39]	yes	22	11	*CCR1*, *CCR2*, *CCRL2*, *CCR3*, *CCR5*, *CCR9*
3	161147744	rs17810546	161.07-161.23	G	0.125	4.56×10^{-18}	9.57×10^{-12}	3.98×10^{-28}	1.36 [1.29-1.44]	yes	22	1	*IL12A*
3	189595248	rs1464510	189.55-189.62	A	0.485	9.49×10^{-24}	3.63×10^{-18}	2.98×10^{-40}	1.29 [1.25-1.34]		22	1	*LPP*
4	123334952	rs13151961	123.19-123.78	G	0.142	6.31×10^{-18}	4.45×10^{-11}	2.18×10^{-27}	0.74 [0.70-0.78]		1	4	*IL2*, *IL21*
6	32713862	rs2187668	gene identified	A	0.258	$<10^{-50}$	$<10^{-50}$	$<10^{-50}$	6.23 [5.95-6.52]	(yes)	1,3	6	*HLA-DQA1*, *HLA-DQB1*
6	138014761	rs2327832	137.92-138.17	G	0.216	1.41×10^{-14}	1.97×10^{-6}	4.46×10^{-19}	1.23 [1.17-1.28]		32	0	*TNFAIP3*
6	159385965	rs1738074	159.24-159.45	A	0.434	3.14×10^{-8}	1.56×10^{-8}	2.94×10^{-15}	1.16 [1.12-1.21]		22	2	*TAGAP*
12	110492139	rs653178	110.19-111.51	G	0.495	6.03×10^{-14}	1.47×10^{-8}	7.15×10^{-21}	1.20 [1.15-1.24]		22	13	*SH2B3*
18	12799340	rs1893217	12.73-12.91	G	0.165	5.52×10^{-7}	1.04×10^{-4}	2.52×10^{-10}	1.17 [1.12-1.23]		17	1	*PTPN2*
New loci, genome-wide significant evidence ($P_{combined}$ <5 x 10^{-8})													
1	2516606	rs3748816	2.40-2.78	G	0.339	4.93×10^{-7}	1.17×10^{-3}	3.28×10^{-9}	0.89 [0.85-0.92]			4	*TNFRSF14*, *MMEL1*
1	25176163	rs10903122	25.11-25.18	A	0.480	3.21×10^{-5}	8.44×10^{-7}	1.73×10^{-10}	0.89 [0.85-0.92]			1	*RUNX3*
1	199158760	rs296547	199.12-199.31	A	0.357	6.46×10^{-5}	1.34×10^{-5}	4.11×10^{-9}	0.89 [0.86-0.92]			2	?
2	68452459	rs17035378[f]	68.39-68.54	G	0.278	1.34×10^{-5}	1.41×10^{-4}	7.79×10^{-9}	0.88 [0.84-0.92]			2	*PLEK*
3	32990473	rs13314993[f]	32.90-33.06	C	0.464	6.87×10^{-6}	1.09×10^{-4}	3.27×10^{-9}	1.13 [1.08-1.17]			2	*CCR4*
3	120601486	rs11712165[f]	120.59-120.78	C	0.394	5.40×10^{-7}	1.72×10^{-3}	8.03×10^{-9}	1.13 [1.08-1.17]			5	*CD80*, *KTELC1*
6	90983333	rs10806425	90.86-91.10	A	0.397	9.46×10^{-6}	9.25×10^{-6}	3.89×10^{-10}	1.13 [1.09-1.17]			1	*BACH2*, *MAP3K7*
6	128320491	rs802734	127.99-128.38	G	0.311	1.36×10^{-6}	1.70×10^{-9}	2.62×10^{-14}	1.17 [1.12-1.22]	yes		2	*PTPRK*, *THEMIS*
8	129333771	rs9792269	129.21-129.37	G	0.238	8.14×10^{-6}	1.00×10^{-4}	3.28×10^{-9}	0.88 [0.84-0.91]			0	?
10	80728033	rs1250552	80.69-80.76	G	0.466	5.80×10^{-8}	1.81×10^{-3}	9.09×10^{-10}	0.89 [0.86-0.92]			1	*ZMIZ1*
11	127886184	rs11221332[f]	127.84-127.99	A	0.237	4.74×10^{-11}	9.98×10^{-7}	5.28×10^{-16}	1.21 [1.16-1.27]	yes		1	*ETS1*
16	11311394	rs12928822	11.22-11.39	A	0.161	1.07×10^{-5}	7.59×10^{-4}	3.12×10^{-8}	0.86 [0.82-0.91]			4	*CIITA*, *SOCS1*, *CLEC16A*
21	44471849	rs4819388	44.42-44.47	A	0.280	3.42×10^{-5}	1.66×10^{-5}	2.46×10^{-9}	0.88 [0.84-0.92]			2	*ICOSLG*
New loci, suggestive evidence (either 10^{-6}>$P_{combined}$>5x10^{-8} and/or P_{GWAS}<10^{-4} and $P_{follow-up}$<0.01)													
1	7969259	rs12727642	7.84-8.13	A	0.185	3.06×10^{-5}	8.21×10^{-4}	9.11×10^{-8}	1.14 [1.09-1.20]			4	*PARK7*, *TNFRSF9*
1	61564451	rs6691768	61.52-61.62	G	0.378	2.63×10^{-5}	1.16×10^{-3}	1.19×10^{-7}	0.90 [0.87-0.94]			1	*NFIA*
1	165678008	rs864537	165.43-165.71	G	0.391	1.01×10^{-7}	9.25×10^{-2}	3.80×10^{-7}	0.91 [0.87-0.94]			1	*CD247*
1	170977623	rs8559637	170.87-171.20	A	0.486	8.15×10^{-5}	5.68×10^{-3}	1.75×10^{-6}	1.10 [1.06-1.14]			1	*FASLG*, *TNFSF18*, *TNFSF4*
3	69335589	rs6806528f	69.27-69.37	A	0.097	4.84×10^{-5}	7.66×10^{-4}	1.46×10^{-7}	1.19 [1.12-1.27]			1	*FRMD4B*
3	170974795	rs10936599	170.84-171.09	A	0.252	2.99×10^{-7}	6.63×10^{-2}	4.57×10^{-7}	1.12 [1.07-1.16]			3	?
6	328546	rs10331809	0.32-0.40	A	0.080	9.14×10^{-6}	1.48×10^{-3}	5.58×10^{-8}	1.21 [1.13-1.29]	yes		1	*IRF4*[g]
7	37341035	rs6974491	37.32-37.41	A	0.170	1.37×10^{-5}	2.63×10^{-3}	1.56×10^{-7}	1.14 [1.09-1.20]			1	*ELMO1*
13	49733716	rs2762051	49.63-49.96	A	0.184	3.35×10^{-5}	5.06×10^{-3}	6.64×10^{-7}	1.13 [1.08-1.18]			0	?
14	68347957	rs4899260	68.24-68.39	A	0.263	4.55×10^{-5}	2.21×10^{-3}	3.92×10^{-7}	1.12 [1.07-1.16]			2	*ZFP36L1*
17	42220599	rs20744044	41.40-42.25	C	0.250	5.03×10^{-5}	5.96×10^{-3}	1.23×10^{-6}	0.90 [0.86-0.94]			10	?
22	20312892	rs2298428	20.14-20.35	A	0.201	2.49×10^{-7}	4.13×10^{-2}	1.84×10^{-7}	1.13 [1.08-1.19]			6	*UBE2L3*, *YDJC*
X	12881445	rs5979785	12.82-12.93	G	0.263	6.32×10^{-6}	2.18×10^{-3}	6.36×10^{-8}	0.88 [0.84-0.92]			1	*TLR7*, *TLR8*

[a]The most significantly associated SNP from each region is shown. [b]LD regions were defined by extending 0.1 cM to the left and right of the focal SNP as defined by the HapMap3 recombination map. All chromosomal in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls. [d]Evidence from positions are based on NCBI build-36 coordinates. [c]Minor allele in all samples logistic regression at a genome-wide significant or suggestive level of significance after conditioning on other associated SNPs (see supplementary table 2). HLA region not tested, but previously known. [e]Selected named genes within or adjacent to the same LD block as the associated SNPs; causality is not proven. In particular, other genes and other causal mechanisms may exist. Gene names underlined are identified from GRAIL[15,16] analysis (see Online Methods) with $P_{text} < 0.01$. [f]These markers were present on the Hap550 but not Hap300 SNP sets, and are not genotyped for 737 cases and 2,596 controls in the stage 1 GWAS, and combined dataset analyses. Only minor changes in P values were observed when these genotypes were imputed and included in analysis. [g]The IRF4 region (specifically rs9738805, $r^2 = 0.08$ with rs10331809 in HapMap CEU) was previously identified as showing strong geographical differentiation[11]. Association with celiac disease was still observed after correction for population stratification using either a structured association approach[34] (corrected $P_{GWAS} = 5.16 \times 10^{-6}$, 478 × 2 × 2 CMH test) or principal components correction (uncorrected $P_{GWAS} = 7.05 \times 10^{-6}$, corrected $P_{GWAS} = 2.28 \times 10^{-5}$, Cochran-Armitage trend tests combined using weighted Z scores; see Online Methods). However, definitive exclusion of population stratification would require family-based association studies.

These regions also contain multiple genes with immunological functions, including *CD247*, *FASLG-TNFSF18-TNFSF4*, *IRF4*, *TLR7-TLR8*, *TNFRSF9* and *YDJC*. Six of the 39 non-HLA regions show evidence for the presence of multiple independently associated variants in a conditional logistic regression analysis (Supplementary Table 2).

We tested the 40 SNPs with the strongest association (Table 2) from each of the known genome-wide significant, new genome-wide significant and new suggestive loci for evidence of heterogeneity across the 12 collections studied. Only the *HLA* region was significant (Breslow-Day test P < 0.05 per 40 tests, rs2187668 $P = 4.8 \times 10^{-8}$), which is consistent with the well-described North-South

Table 3 Celiac risk variants correlated with cis gene expression

SNP[a]	Chr	SNP position[b]	Probe Centre Position[b]	Illumina ArrayAddressID	Expression dataset[c]	Gene name	eQTL P value[d]
Loci with genome-wide significant evidence ($P_{combined}$ <5 x 10 $^{-8}$)							
rs3748816	1	2516606	2412221	650452	HT-12	*PLCH2*	1.66 x 10 $^{-5}$
rs3748816	1	2516606	2482955	6520725	Ref-8v2 + HT-12	*TNFRSF14*	1.30 x 10 $^{-3}$
rs3748816	1	2516606	2510429	6250338	Ref-8v2	*C1orf93*	1.16 x 10 $^{-4}$
rs3748816	1	2516606	2533115	2070246	Ref-8v2 + HT-12	*MMEL1*	1.03 x 10 $^{-20}$
rs296547	1	199158760	198880146	1300279	Ref-8v2 + HT-12	*DDX59*	2.45 x 10 $^{-5}$
rs842647	2	60972975	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	3.30 x 10 $^{-10}$
rs13003464[e]	2	61040333	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	6.39 x 10 $^{-11}$
rs3816281[f]	2	68461451	68461957	4810020	Ref-8v2 + HT-12	*PLEK*	7.97 x 10 $^{-26}$
rs917997	2	102437000	102418571	6520180	Ref-8v2 + HT-12	*IL18RAP*	7.35 x 10 $^{-87}$
rs13010713	2	181704290	181593865	1780433	HT-12	*UBE2E3*	4.93 x 10 $^{-5}$
rs13098911	3	46210205	45964449	6550333	Ref-8v2 + HT-12	*CXCR6*	9.66 x 10 $^{-6}$
rs13098911	3	46210205	46255176[g]	2190671	HT-12	*CCR3*	5.50 x 10 $^{-10}$
rs13098911	3	46210205	46255176[g]	7570670	Ref-8v2	*CCR3*	5.69 x 10 $^{-4}$
rs6441961[d]	3	46327388	46255176[h]	2190671	HT-12	*CCR3*	2.87 x 10 $^{-19}$
rs6441961[d]	3	46327388	46255176[h]	7570670	Ref-8v2	*CCR3*	1.02 x 10 $^{-4}$
rs11922594[f]	3	120608512	120683364[i]	6550288	Ref-8v2 + HT-12	*KTELC1*	5.09 x 10 $^{-17}$
rs11922594[f]	3	120608512	120683364[i]	3850161	Ref-8v2 + HT-12	*KTELC1*	7.34 x 10 $^{-6}$
rs10806425	6	90983333	90878075	3520349	HT-12	*BACH2*	1.92 x 10 $^{-3}$
rs1738074	6	159385965	159380068	5890739	Ref-8v2 + HT-12	*TAGAP*	1.99 x 10 $^{-3}$
rs1738074	6	159385965	159381094[j]	5360364	HT-12	*TAGAP*	3.23 x 10 $^{-4}$
rs1738074	6	159385965	159381094[j]	4860242	HT-12	*TAGAP*	2.18 x 10 $^{-3}$
rs1250552	10	80728033	80622540	2450131	Ref-8v2 + HT-12	*ZMIZ1*	1.80 x 10 $^{-3}$
rs653178	12	110492139	110399552	6560301	Ref-8v2 + HT-12	*SH2B3*	9.24 x 10 $^{-12}$
rs653178	12	110492139	110710447	840253	Ref-8v2 + HT-12	*ALDH2*	1.44 x 10 $^{-4}$
rs653178	12	110492139	110894406[k]	2070736	HT-12	*TMEM116*	3.68 x 10 $^{-4}$
rs653178	12	110492139	110894406[k]	3190129	Ref-8v2	*TMEM116*	1.51 x 10 $^{-3}$
rs12928822	16	11311394	11335627	4540072	Ref-8v2 + HT-12	*C16orf75*	1.02 x 10 $^{-8}$
rs4819388	21	44471849	44049567	7200373	Ref-8v2	*RRP1*	2.62 x 10 $^{-3}$
Loci with suggestive evidence (either A. 10 $^{-6}$>$P_{combined}$>5x10 $^{-8}$ and/or B. P_{GWAS} <10-4 and $P_{follow-up}$<0.01)							
rs12726642	1	7969259	7956138	610193	Ref-8v2 + HT-12	*PARK7*	9.76 x 10 $^{-15}$
rs864537	1	165678008	165710482[l]	6290400	Ref-8v2 + HT-12	*CD247*	1.77 x 10 $^{-9}$
rs864537	1	165678008	165710482[l]	3890689	HT-12	*CD247*	2.93 x 10 $^{-7}$
rs6974491	7	37341035	37157761	2750154	Ref-8v2 + HT-12	*ELMO1*	5.40 x 10 $^{-6}$
rs2074404	17	42220599	41824345	3520672	Ref-8v2 + HT-12	*LRRC37A*	1.17 x 10 $^{-4}$
rs2074404	17	42220599	42106695[m]	5260138	Ref-8v2 + HT-12	*NSF*	1.20 x 10 $^{-5}$
rs2074404	17	42220599	42106695[m]	1410484	HT-12	*NSF*	4.28 x 10 $^{-4}$
rs2074404	17	42220599	42223012	4070615	HT-12	*WNT3*	2.77 x 10 $^{-3}$
rs2074404	17	42220599	42485154	4880037	HT-12	*LOC388397*	1.78 x 10 $^{-9}$
rs2298428	22	20312892	20308188	1230242	Ref-8v2 + HT-12	*UBE2L3*	1.96 x 10 $^{-90}$
rs5979785	X	12881445	12842944[n]	6480360	Ref-8v2 + HT-12	*TLR8*	3.88 x 10 $^{-13}$
rs5979785	X	12881445	12842944[n]	3390612	Ref-8v2 + HT-12	*TLR8*	1.07 x 10 $^{-7}$

See supplementary Figures 2 and 3 for detailed results and supplementary table 3 for more details of Illumina expression probes.
[a]We tested the SNP with the strongest association from 34 of 39 non-HLA loci ($P_{combined}$ < 10^{-6}, table 2), Hap300 proxy SNPs for 4 further loci, and a second independently associated SNP from 6 loci, for correlation with gene expression in PAXgene blood RNA in up to 1,349 individuals. One locus (containing *ETS1*) where an adequate proxy SNP was not available was not included for the eQTL analysis. SNP-gene expression correlations were tested for probes within a 1-Mb window. Results are presented for SNPs showing significant correlations with cis gene expression after controlling false-discovery rate at 5% (corresponding to P < 0.0028). [b]All chromosomal positions are based on NCBI build-36 coordinates. Probe center position was determined by re-mapping probe sequences to the human transcriptome and calculated from the midpoint of the transcript start and transcript end positions in genomic coordinates. [c]'HT-12' comprise 1,240 individuals with blood gene expression assayed using Illumina Human HT-12v3 arrays; 'Ref-8v2' comprise 229 individuals with blood gene expression assayed using Illumina Human-Ref-8v2 arrays (see Online Methods). [d]Spearman rank correlation of genotype and residual variance in transcript expression. Meta-analysis eQTL P value shown if both datasets had identical probes. [e]Second, independently associated SNP from this locus. [f]Proxy SNP, r^2 = 0.61 in HapMap CEU with most associated SNP rs11712165. [g–n]Different Illumina probe sequences with the same probe center position.

gradient in *HLA* allele frequency in European populations, and more specifically for *HLA-DQ* in celiac disease[14].

We observed no evidence for interaction between each of the 26 genome-wide significant non-HLA loci, which is consistent with what has been reported for other complex diseases so far. However, we did observe weak evidence for lower effect sizes at non-HLA loci in high risk *HLA-DQ2.5cis* homozygotes, similar to what has been observed in type 1 diabetes[7].

To obtain more insight into the functional relatedness of the celiac disease risk loci, we applied GRAIL, a statistical tool that uses text mining of PubMed abstracts to annotate candidate genes from loci associated with common disease risk[15,16]. To assess the sensitivity of this tool (using known loci as a positive control), we first performed a 'leave-one-out' analysis of the 27 genome-wide significant celiac disease loci (including *HLA-DQ*). GRAIL scores of $P_{text} < 0.01$ were obtained for 12 of the 27 loci (44% sensitivity; Table 2). Factors that limit the sensitivity of GRAIL include biological pathways being both known (a 2006 dataset is used to avoid GWAS-era studies) and published in the literature. We then applied GRAIL analysis, using the 27 known regions as a seed, to all 49 regions (49 SNPs) with $10^{-3} > P_{combined} > 5 \times 10^{-8}$ and obtained GRAIL $P_{text} < 0.01$ for 9 regions (18.4%). As a control, only 5.5% (279 of 5,033) of randomly selected Hap550 SNPs reached this threshold. In addition to the five 'suggestive' loci shown in Table 2, GRAIL annotated four further interesting gene regions with lower significance in the combined association results: rs944141-*PDCD1LG2* ($P_{combined} = 4.4 \times 10^{-6}$), rs976881-*TNFRSF8* ($P_{combined} = 2.1 \times 10^{-4}$), rs4682103-*CD200- BTLA* ($P_{combined} = 6.8 \times 10^{-6}$) and rs4919611-*NFKB2* ($P_{combined} = 6.1 \times 10^{-5}$). There appeared to be further enrichment for genes of immunological interest that are not GRAIL-annotated in the $10^{-3} > P_{combined} > 5 \times 10^{-8}$ significance window, including rs3828599-*TNIP1* ($P_{combined} = 1.55 \times 10^{-4}$), rs8027604-*PTPN9* ($P_{combined} = 1.4 \times 10^{-6}$) and rs944141-*CD274* ($P_{combined} = 4.4 \times 10^{-6}$). Some of these findings, for which neither genome-wide significant nor suggestive association is achieved, are likely to comprise part of a longer tail of disease-predisposing common variants with weaker effect sizes. Definitive assessment of these biologically plausible regions would require genotyping and association studies using much larger sample collections than the present study.

We previously showed that there is considerable overlap between risk loci for celiac disease and type 1 diabetes[17], as well as between risk loci for celiac disease and rheumatoid arthritis[18], and more generally, there is now substantial evidence for shared risk loci between the common chronic immune-mediated diseases[6]. To update these observations, we searched 'A Catalog of Published Genome Wide Association Studies' (accessed 18 November 2009)[19] and the HuGE database[20]. We found some evidence (requiring a published association report of $P < 1 \times 10^{-5}$) of shared loci with at least one other inflammatory or immune-mediated disease for 18 of the current 27 genome-wide significant celiac disease risk regions. We defined shared regions as the broad linkage disequilibrium block; however, different SNPs are often reported in different diseases, and at only 3 of the 18 shared regions are associations across all diseases with the same SNP or a proxy SNP in $r^2 > 0.8$ in HapMap CEU. Currently, nine regions seem to be specific to celiac disease and might reflect distinctive disease biology, including the regions containing rs296547 and rs9792269 and the regions around *CCR4, CD80, ITGA4, LPP, PLEK, RUNX3* and *THEMIS*. In fact, locus sharing between diseases is probably greater because of both stochastic variation in results from sample size limitations and regions that have a genuinely stronger effect size in one disease and weaker effect size in another.

Genetic variation in *ETS1* has recently been reported to be associated with systemic lupus erythematosus (SLE) in the Chinese population, although it is not associated with SLE in European populations[21]. The most

strongly associated celiac disease (European population) SNP, rs11221332, and the most strongly associated SLE (Chinese population) SNP, rs6590330, map 70 kb apart. Inspection of the HapMap phase II data shows broadly similar linkage disequilibrium patterns between Chinese (CHB) and European (CEU) populations in this region, with the two associated SNPs in separate nonadjacent linkage disequilibrium blocks. Thus, distinct common variants within the same gene can predispose to different autoimmune diseases across different ethnic groups.

Exploring the function of celiac disease risk variants

Celiac disease risk variants in the HLA genes alter protein structure and function[4]. However, we identified only four nonsynonymous SNPs with evidence for association with celiac disease ($P_{GWAS} < 10^{-4}$) from the other 26 genome-wide significant associated regions (rs3748816-*MMEL1*, rs3816281-*PLEK*, rs196432-*RUNX3*, rs3184504-*SH2B3*). Although comprehensive regional resequencing is required to test the possibility that coding variants contribute to the observed association signals, more subtle effects of genetic variation on gene expression are the more likely functional mechanism for complex disease genes. With this in mind, we performed a meta-analysis of new and published genome-wide eQTL datasets comprising 1,469 human whole blood (PAXgene) samples reflecting primary leukocyte gene expression. We applied a new method, transcriptional components, to remove a substantial proportion of inter-individual nongenetic expression variation and performed eQTL meta-analysis on the residual expression variation (Online Methods).

We assessed 38 of the 39 genome-wide significant and suggestive celiac disease-associated non-HLA loci (Table 2) for *cis* expression-genotype correlations. We tested the SNP with the strongest association from each region. However, for five regions the most associated SNP was not genotyped in the eQTL samples (Hap300 data); instead, for four of these, we tested a proxy SNP ($r^2 > 0.5$ in HapMap CEU). In addition, for six loci showing evidence of multiple independent associations in conditional regression analyses, we tested a second SNP that showed independent association with celiac disease for eQTL analysis. In total, we assessed 44 independent non-HLA SNP associations in peripheral whole blood samples genotyped on the Illumina Hap300 BeadChip and either Illumina Ref8 or HT12 expression arrays, correlating each SNP with data from gene probes mapping within a 1-Mb window.

We identified significant (Spearman $P < 0.0028$, corresponding to 5% false-discovery rate) eQTLs at 20 of 38 (52.6%) non-HLA celiac loci tested (Table3 and Supplementary Figs. 2 and 3). Some loci had evidence of eQTLs with multiple probes, genes or SNPs (Table 3). We assessed whether the number of SNPs with *cis*-eQTL effects out of the 44 SNPs that we tested was significantly higher than expected. On average, eQTL SNPs had a substantially higher minor allele frequency (MAF) than non-eQTL SNPs in the 294,767 SNPs tested. To correct for this, we selected 44 random SNPs that had an equal MAF distribution and determined for how many of these MAF-matched SNPs eQTLs were observed. There were a significantly higher number of eQTL SNPs ($P = 9.3 \times 10^{-5}$, 10^6 permutations) among the celiac disease–associated SNPs than expected by chance (22 observed eQTL SNPs versus 7.8 expected eQTL SNPs). Therefore, the celiac disease–associated regions are greatly enriched for eQTLs. These data indicate that some risk variants might influence celiac disease susceptibility through a mechanism of altered gene expression. Candidate genes with a significant eQTL where the peak eQTL signal and peak case- control association signal are similar (Supplementary Fig. 3) include *MMEL1*, *NSF*, *PARK7*, *PLEK*, *TAGAP*, *RRP1*, *UBE2L3* and *ZMIZ1*.

We also assessed the coexpression of genes that mapped within 500 kb of SNPs that showed the strongest case-control association

from the 40 genome-wide significant and suggestive celiac disease loci in an analysis of the 33,109 human Affymetrix Gene Expression Omnibus dataset. This analysis loses power to detect tissue-specific correlations from the use of numerous tissue types, but it greatly gains power from the large sample size. We detected several distinct coexpression clusters (Pearson correlation coefficient between genes >0.5), including four clusters of immune-related genes that contain at least one gene from 37 of the 40 genome-wide significant and suggestive loci (Fig. 1). These data further demonstrate that genes from celiac disease risk loci map to multiple distinct immunological pathways involved in disease pathogenesis.

DISCUSSION

We previously reported that most celiac genetic risk variants mapped near genes that are functional in the immune system[22], and this remains true for the 13 new genome-wide significant and 13 new suggestive risk variants from the current study. We can now refine these observations and highlight specific immunological pathways that are relevant to the pathogenesis of celiac disease.

One key pathway worth highlighting is T-cell development in the thymus. The rs802734 linkage disequilibrium block contains the recently identified gene *THEMIS* (thymus-expressed molecule involved in selection). *THEMIS* has a key regulatory role in both

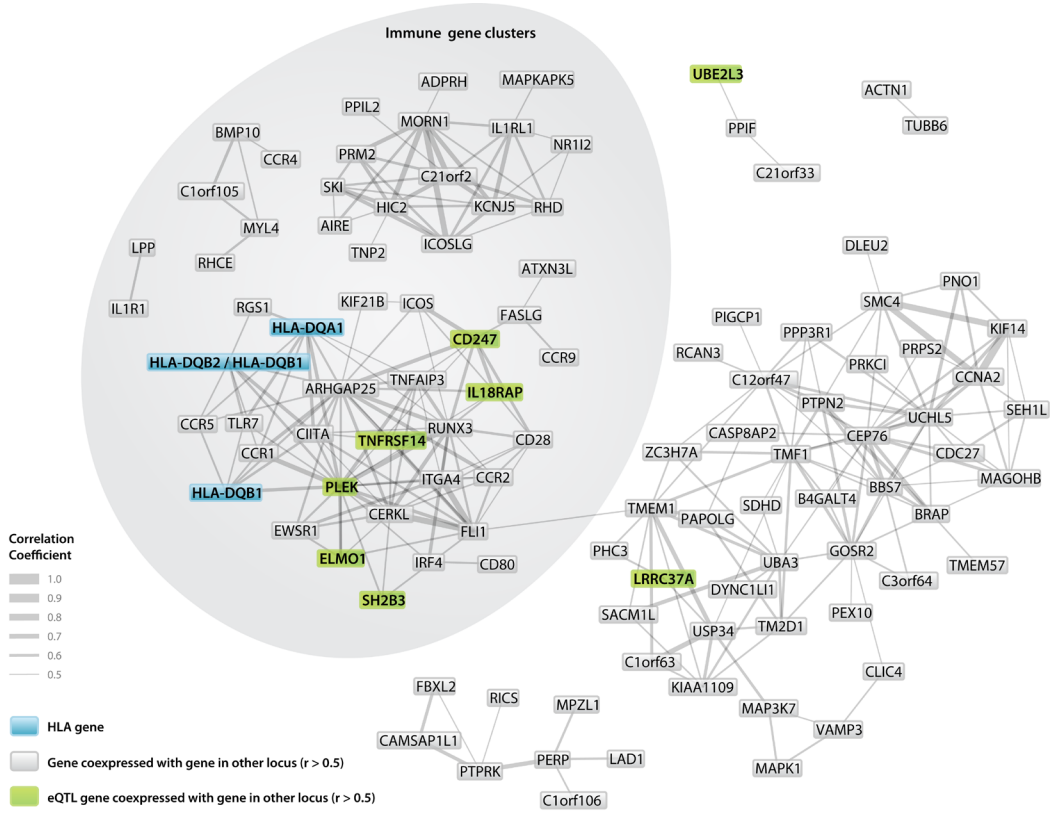


Figure 1 Coexpression analysis of genes mapping to 40 genome-wide significant and suggestive celiac disease regions in 33,109 heterogenous human samples from the Gene Expression Omnibus. Genes mapping within a 1-Mb window of associated SNPs (table 2) were tested for interaction with genes from other loci. Interactions with Pearson correlation > 0.5 are shown ($P < 10^{-100}$). Only the genes known to contain causal mutations (HLA-DQA1, HLA-DQB1) were analyzed from the HLA region; HLA-DQB2/HLA-DQB1 is a single expression probeset mapping to both genes. No probe for THEMIS was present on the earlier version of the U133 array; however, in a subset analysis of U133 Plus2.0 data, THEMIS is coexpressed in the major immune gene cluster.

positive and negative T-cell selection during late thymocyte development[23]. Furthermore, the rs10903122 linkage disequilibrium block contains *RUNX3*, a master regulator of $CD8^+$ T lymphocyte development in the thymus[24,25]. *TNFRSF14* (LIGHTR, rs3748816 linkage disequilibrium block) has widespread functions in peripheral leukocytes and a crucial role in promoting thymocyte apoptosis[26]. The *ETS1* transcription factor (rs11221332 linkage disequilibrium block) is also active in peripheral leukocytes; however, it is also a key player in thymic $CD8^+$ lineage differentiation, acting in part by promoting *RUNX3* expression[27].

The importance of the thymus in the pathogenesis of autoimmune diseases has been previously emphasized by the established role of thymectomy in the treatment of myasthenia gravis. In type 1 diabetes, disease-associated genetic variation in the insulin gene INS causes altered thymic insulin expression and subsequent T-cell tolerance for insulin as a self-protein[28]. However, the importance of thymic T-cell regulation in the etiology of celiac disease has not been previously recognized. It is conceivable that the associated variants might alter biological processes before thymic MHC-ligand interactions. Alternatively, it is now clear that exogenous antigen presentation and selection occurs in the thymus through migratory dendritic cells; this has been demonstrated for skin and has been hypothesized for food antigens[29,30]. These findings suggest that it would be worthwhile to investigate immunological and pharmacological modifiers of T-cell tolerance more generally in autoimmune diseases.

A second pathway worth noting is the innate immune detection of viral RNA. Although the association signal at rs5979785 ($P_{combined} = 6.36 \times 10^{-8}$) in the *TLR7-TLR8* region is just outside our genome-wide significance threshold, we observe a strong effect of rs5979785 on *TLR8* expression in whole blood. Both TLRs recognize viral RNA. Taken together with the recent observation that rare loss-of-function mutations in the enteroviral response gene *IFIH1* are protective against type 1 diabetes[31], these findings implicate viral infection (and the nature of the host response to infection) as a putative environmental trigger that could be common to these autoimmune diseases.

A third pathway involves T- and B-cell co-stimulation (or co-inhibition). This class of molecules controls the strength and nature of the response to T-cell or B-cell (immunoglobulin) receptor activation by antigens. We observe multiple regions with genes (*CTLA4-ICOS-CD28, TNFRSF14, CD80, ICOSLG, TNFRSF9, TNFSF4*) from this class of ligand-receptor pairs, indicating that fine control of the adaptive immune response might be altered in individuals at risk of celiac disease.

A final pathway involves cytokines, chemokines and their receptors. Our previous report discussed the function of the 2q11–12 interleukin receptor cluster (*IL18RAP* and so on), the 3p21 chemokine receptor cluster (*CCR5* and so on) and the loci containing *IL2-IL21* and *IL12A*[22]. We now report additional loci containing *TNFSF18* and *CCR4*.

We estimate that the current celiac disease variants, including the major celiac disease-associated HLA variant, *HLA-DQ2.5cis*, less common celiac disease–associated haplotypes in the HLA (*HLA-DQ8; HLA-DQ2.5trans; HLADQ2.2*), and the additional 26 definitively implicated loci explain about 20% of total celiac disease variance, which would represent 40% of genetic variance, assuming a heritability of 0.5. A long tail of common variants with low effect size, along with highly penetrant rare variants (both at the established loci and elsewhere in the genome), might contribute substantially to the remaining heritability.

We observed different haplotypes within the *ETS1* region associated with celiac disease in Europeans and SLE in the Chinese population. For some autoimmune diseases studied in European origin populations, although the same linkage disequilibrium block has been associated, the association is with a different

haplotype. In some cases, the same variants are associated, but the direction of association is opposite (for example, rs917997-*IL18RAP* in celiac disease versus type 1 diabetes). We believe further exploration of these signals might reveal critical differences in the nature of the immune system perturbation between these diseases.

Previously, investigators have observed that only a small proportion of GWAS signals involve coding variants and have suggested that these variants might instead influence regulation of gene expression. Here we show that over half the variants associated with celiac disease are correlated with expression changes in nearby genes. This mechanism is likely to explain the function of some risk variants for other common, complex diseases. Further research is needed to definitively determine at each locus both the variants that can cause celiac disease and their functional mechanisms.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession numbers. Expression data are available in GEO (http:// www.ncbi.nlm.nih.gov/geo/) as GSE20142 and GSE20332.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank Coeliac UK for assistance with direct recruitment of individuals with celiac disease, and UK clinicians (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis and K. Moriarty) who recruited the celiac disease blood samples described in our previous studies[1,22]. We thank the genotyping facility of the UMCG (J. Smolonska and P. van der Vlies) for generating part of the GWAS and replication data and the gene expression data; R. Booij and M. Weenstra for preparation of Italian samples; H. Ahola, A. Heimonen, L. Koskinen, E. Einarsdottir and K. Löytynoja for their work on Finnish sample collection, preparation and data handling; and E. Szathmári, J.B.Kovács, M. Lörincz and A. Nagy for their work with the Hungarian families. The Health2000 organization, Finrisk consortium, K. Mustalahti, M. Perola, K. Kristiansson and J. Koskinen are thanked for providing the Finnish control genotypes. We thank D.G. Clayton and N. Walker for providing T1DGC data in the required format. We thank the Irish Transfusion Service and Trinity College Dublin Biobank for control samples and V. Trimble, E. Close, G. Lawlor, A. Ryan, M. Abuzakouk, C. O'Morain and G. Horgan for celiac disease sample collection and preparation We acknowledge DNA provided by Mayo Clinic Rochester and thank M. Bonamico and M. Barbato (Department of Paediatrics, Sapienza University of Rome, Italy) for recruiting individuals. We thank Polish clinicians for recruitment of individuals with celiac disease (Z. Domagala, A. Szaflarska-Poplawska, B. Oralewska, W. Cichy, B. Korczowski, K. Fryderek, E. Hapyn, K. Karczewska, A. Zalewska, I. Sakowska-Maliszewska, R. Mozrzymas, A. Zabka, M. Kolasa and B. Iwanczak). We thank M. Szperl for isolating DNA from blood samples provided by the Children's Memorial Health Institute (Warsaw, Poland). Dutch and UK genotyping for the second celiac disease GWAS was funded by the Wellcome Trust (084743 to D.A.v.H.). Italian genotyping for the second celiac disease GWAS was funded by the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009 to C.W.) and by the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.). E.G. is funded by the Italian Ministry of Healthy (grant RC2009). L.H.v.d.B. acknowledges funding from the Prinses Beatrix Fonds, the Adessium foundation and the Amyotrophic Lateral Sclerosis Association. L.F. received a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031) and a VENI grant from NWO (ZonMW grant 916.10.135). P.C.A.D. is an MRC Clinical Training Fellow (G0700545). G.T. received a Ter Meulen Fund grant from the Royal Netherlands Academy of Arts and Sciences (KNAW). The gene expression study was funded in part by COPACETIC (EU grant 201379). This study makes use of data generated by the Wellcome Trust Case-Control Consortium 2 (WTCCC2). A full list of the WTCCC2 investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the WTCCC2 project was provided by the Wellcome Trust under award 085475. This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD) and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418.

We acknowledge the use of BRC Core Facilities provided by the financial support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. We acknowledge funding from the NIH: DK050678 and DK081645 (to S.L.N.), NS058980 (to R.A.O.); and DK57892 and DK071003 (to J.A.M.). The collection of Finnish and Hungarian subjects with celiac disease was funded by the EU Commission (MEXT-CT-2005-025270),

the Academy of Finland, Hungarian Scientific Research Fund (contract OTKA 61868), the University of Helsinki Funds, the Competitive Research Funding of the Tampere University Hospital, the Foundation of Pediatric Research, the Sigrid Juselius Foundation and the Hungarian Academy of Sciences (2006TKI247 to R.A.). Funding for the collection and genotyping of the Polish samples was provided by UMC Cooperation Project (6/06/2006/NDON). R.M. is funded by Science Foundation Ireland. C. Núñez has a FIS contract (CP08/0213). The Dublin Centre for Clinical Research contributed to collection of samples from affected individuals and is funded by the Irish Health Research Board and the Wellcome Trust. Finally, we thank all individuals with celiac disease and control individuals for participating in this study.

AUTHOR CONTRIBUTIONS

D.A.v.H. and C.W. designed, co-ordinated and led the study. Experiments were performed in the labs of C.W., D.A.v.H., C.A.M., P.D. and P.M.G. Major contributions were: (i) DNA sample preparation: P.C.A.D., G.T., K.A.H., J.R., A.Z. and P.S.; (ii) genotyping: P.C.A.D., G.T., K.A.H., A.C., J.R. and R.G.; (iii) expression data generation: H.J.M.G., L.H.v.d.B., R.A.O., R.K.W. and L.F.; (iv) case-control association analyses: P.C.A.D., G.T., L.F., J.C.B. and D.A.v.H.; (v) expression analyses: L.F., G.A.R.H. and R.S.N.F.; (vi) manuscript preparation: P.C.A.D., G.T., L.F., R.S.N.F., G.A.R.H., J.C.B., C.W. and D.A.v.H. Other authors contributed variously to sample collection and all other aspects of the study. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests. Published online at http://www.nature.com/naturegenetics/.

Reprints and permissions information is available online at http://npg.nature.com/ reprintsandpermissions/.

REFERENCES

1. van Heel, D.A. et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827–829 (2007).

2. van Heel, D.A. & West, J. Recent advances in coeliac disease. Gut 55, 1037–1046 (2006).

3. Sollid, L.M. et al. Evidence for a primary association of celiac disease to a particular HLA-DQ α/β heterodimer. J. Exp. Med. 169, 345–350 (1989).

4. Kim, C.Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L.M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc. Natl. Acad. Sci. USA 101, 4175–4179 (2004).

5. Henderson, K.N. et al. A structural and immunological basis for the role of human leukocyte antigen DQ8 in celiac disease. Immunity 27, 23–34 (2007).

6. Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet. 10, 43–55 (2009).

7. Barrett, J.C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41, 703–707 (2009).

8. Barrett, J.C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 40, 955–962 (2008).

9. Anderson, C.A. et al. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am. J. Hum. Genet. 83, 112–119 (2008).

10. Jacobs, K.B. et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat. Genet. 41, 1253–1257 (2009).

11. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

12. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. 32, 381–385 (2008).

13. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. 32, 227–234 (2008).

14.Karell, K. et al. HLA types in celiac disease patients not carrying the DQA1*05–DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum. Immunol. 64, 469–477 (2003).

15. Raychaudhuri, S. et al. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet. 41, 1313–1318 (2009).

16. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 5, e1000534 (2009).

17. Smyth, D.J. et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med. 359, 2767–2777 (2008).

18.Coenen, M.J. et al. Common and different genetic background for rheumatoid arthritis and coeliac disease. Hum. Mol. Genet. 18, 4195–4203 (2009).

19. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106, 9362–9367 (2009).

20.Yu, W., Clyne, M., Khoury, M.J. & Gwinn, M. Phenopedia and Genopedia: Disease- centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics 26, 145–146 (2010).

21. Han, J.W. et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. 41, 1234–1237 (2009).

22. Hunt, K.A. et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40, 395–402 (2008).

23. Allen, P.M. Themis imposes new law and order on positive selection. Nat. Immunol. 10, 805–806 (2009).

24. Sato, T. et al. Dual functions of Runx proteins for reactivating CD8 and silencing CD4 at the commitment process into CD8 thymocytes. Immunity 22, 317–328 (2005).

25. Woolf, E. et al. Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. Proc. Natl. Acad. Sci. USA 100, 7731–7736 (2003).

26. Wang, J. & Fu, Y.X. LIGHT (a cellular ligand for herpes virus entry mediator and lymphotoxin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC/self-peptide. J. Immunol. 170, 3986–3993 (2003).
27. Zamisch, M. et al. The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8 T cell differentiation in the thymus. J. Exp. Med. 206, 2685–2699 (2009).
28. Vafiadis, P. et al. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. Nat. Genet. 15, 289–292 (1997).
29. Bonasio, R. et al. Clonal deletion of thymocytes by circulating dendritic cells homing to the thymus. Nat. Immunol. 7, 1092–1100 (2006).
30. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. Nat. Rev. Immunol. 9, 833–844 (2009).
31. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324, 387–389 (2009).
32. Trynka, G. et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-κB signalling. Gut 58, 1078–1083 (2009).
33. Garner, C.P. et al. Replication of celiac disease UK genome-wide association study results in a US population. Hum. Mol. Genet. 18, 4219–4225 (2009).
34. Plenge, R.M. et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat. Genet. 39, 1477–1482 (2007).
35. Franke, L. et al. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am. J. Hum. Genet. 82, 1316–1333 (2008).
36.Purcell, S. et al. PLINK: a tool set for whole-genome association and population- based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
37. Price, A.L. et al. Principal components analysis corrects for stratification in genome- wide association studies. Nat. Genet. 38, 904–909 (2006).
38. Yu, K. et al. Population substructure and control selection in genome-wide association studies. PLoS One 3, e2551 (2008).
39. Risch, N.J. Searching for genetic determinants in the new millennium. Nature 405, 847–856 (2000).
40.Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210 (2002).
41. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193 (2003).
42. Sherlock, G. Analysis of large-scale gene expression data. Brief. Bioinform. 2, 350–362 (2001).
43. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA 97, 10101–10106 (2000).
44. Heap, G.A. et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. Hum. Mol. Genet. 19, 122–134 (2010).
45.Heap, G.A. et al. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med. Genomics 2, 1 (2009).
46. Franke, L. & Jansen, R.C. eQTL analysis in humans. Methods Mol. Biol. 573, 311–328 (2009).

ONLINE METHODS

Subjects. Written informed consent was obtained from all subjects, with Ethics Committee/Institutional Review Board approval. All individuals are of European ancestry. Affected celiac individuals were diagnosed according to standard clinical, serological and histopathological criteria, including small intestinal biopsy. DNA samples were from blood, lymphoblastoid cell lines or saliva. A more detailed description of subjects is provided in a Supplementary Note.

GWAS genotyping. For an overview, see Table 1. UK(1) case and control genotyping has been described[1,7]. Illumina 670-Quad and 1.2M-Duo (custom chips designed for the WTCCC2 and comprising Hap550/1M and common CNV content) and 610-Quad genotyping was performed in London, Hinxton and Groningen. Bead intensity data was normalized for each sample in BeadStudio, R and theta values exported and genotype calling performed using a custom algorithm[1,35]. A detailed description of genotype calling steps is provided in a Supplementary Note.

Quality control steps were performed in the following order. First, very low call rate samples and SNPs were excluded. SNPs were excluded from all sample collections if any collection showed call rates <95% or deviation from Hardy-Weinberg equilibrium ($P < 0.0001$) in controls. Samples were excluded for call rate <98%, incompatible recorded gender and genotype-inferred gender, ethnic outliers (identified by multi-dimensional scaling plots of samples merged with HapMap Phase II data), duplicates and first-degree relatives. We excluded 22 of 417 SNPs showing apparent association ($P_{GWAS} < 10^{-4}$) after visual inspection of R theta plots suggested possible bias.

The over-dispersion factor of association test statistics (genomic control inflation factor), λ_{GC}, was calculated using observed versus expected values for all SNPs in PLINK.

Follow-up genotyping. For an overview, see Table 1. Finnish controls (12) were genotyped on the 610-Quad BeadChip; other samples were genotyped using Illumina GoldenGate BeadXpress assays in London and Groningen. Genotyping calling was performed in BeadStudio for combined cases and controls in each separate collection, with the exception of the Finnish collection, and whole genome amplified samples (89 Irish cases and 106 Spanish controls). Quality control steps were performed as for the GWAS. In total, 131 of 144 SNPs passed quality control and visual inspection of genotype clouds.

SNP association analysis. Analyses were performed using PLINK v1.07[36], mostly using the Cochran-Mantel-Haenzel test. Logistic regression analyses were used to define the independence of association signals within the same linkage disequilibrium block, with group membership included as a factorized covariate.

Genotype imputation was performed for samples genotyped on the Hap300 using BEAGLE and CEU, TSI, MEX and GIH reference samples from HapMap3. Association analysis was performed using logistic regression on posterior genotype probabilities, with group membership included as a factorized covariate.

Structured association tests were performed using PLINK as described using genetically matched cases and controls within collections identified by identity by state similarity across autosomal non-HLA SNPs[34] (settings–ppc 0.001–cc, clusters constrained by the five collections). Principal components analysis was performed using EIGENSTRAT and a set of 12,810 autosomal non-HLA SNPs chosen for low LD and ancestry information[37,38]; association tests were corrected for the top 10 principal components and combined using weighted Z scores.

The fraction of additive variance was calculated using a liability threshold model[39] assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the combined replication panel. Genetic variance was calculated assuming 50% heritability.

GRAIL analysis. We performed GRAIL analysis (http://www.broadinstitute. org/mpg/grail/grail.php) using HG18 and Dec2006 PubMed datasets, default settings for SNP rs number submission, and the 27 genome-wide significant celiac disease risk loci (most associated SNP) as seeds. As a query, we used either associated SNPs or 101 × 50 randomly chosen Hap550 SNP datasets (5,050 SNPs, of which 5,033 mapped to the GRAIL database).

Identification of transcriptional components. We noted that the power of eQTL studies in humans is limited by substantial observed interindividual variation in expression measurements due to nongenetic factors, and therefore developed a method, 'transcriptional components', to remove a large component of this variation (manuscript in preparation). Expression data from 42,349 heterogeneous human samples hybridized to Affymetrix HG-U133A (GEO accession number: GPL96) or HG-U133 Plus 2.0 (GEO accession number: GPL570) Genechips were downloaded[40]. Samples missing data for >150 probes were excluded, and only probes available on both platforms were analyzed, resulting in expression data for 22,106 probes and 41,408 samples. We performed quantile normalization using the median rank distribution[41] and $\log_2$ transformed the data, ensuring an identical distribution of expression signals for every sample, discarding previous normalization and transformation steps.

Initial quality control (QC) was performed by applying principal component analysis (PCA) on the sample correlation matrix (pair-wise Pearson correlation coefficients between all

samples). The first principal component (PC), explaining ~80–90% of the total variance[42,43], describes probe-specific variance. 6,375 samples with correlation R < 0.75 of the sample array with this PC were considered outliers of lesser quality and excluded from analysis. We excluded entire GEO datasets where >25% of the samples were outliers (prob- ably expression ratios versus a reference, not absolute data). The final dataset comprised 33,109 samples (17,568 GPL96 and 15,541 GPL570 samples), and we repeated the normalization and transformation on the originally deposited expression values of these post-quality control samples.

We next applied PCA on the pairwise 22,106 × 22,106 probe Pearson correlation coefficient matrix assayed on the 33,109 sample dataset (our fast C++ tool, MATool, is available upon request), attempting to simplify the structure of the data. Here, PCA represents a transformation of a set of correlated probes into sets of uncorrelated linear additions of probe expression signals (eigen- vectors) that we name transcriptional components (TCs). Each TC is a weighted sum of probe expression signals and eigenvector probe coefficients. These TC scores can be calculated for each observed expression array sample (reflecting the TC activity per sample).

Subjects for expression-genotype correlation. We obtained peripheral blood DNA and RNA (PAXgene) from Dutch and UK individuals who were disease cases or controls for GWAS studies (Supplementary Table 1). All samples had been genotyped for a common SNP set on Illumina platforms. Analysis was confined to 294,767 SNPs that had a MAF ≥ 5%, call-rate ≥ 95% and exact HWE P > 0.001. RNA from the samples was hybridized to either Illumina HumanRef-8 v2 arrays (229 samples, Ref-8v2) or Illumina HumanHT-12 arrays (1,240 samples, HT-12), and raw probe intensity extracted using BeadStudio. The Ref-8v2 samples were jointly quantile normalized and $\log_2$ transformed, as were the HT-12 samples. Subsequent analyses were also conducted separately for these datasets, up to the eventual eQTL mapping, which uses a meta-analysis framework, combining eQTL results from both arrays. HT-12 and Ref-8v2 arrays are different, but share many probes with identical probe sequences. Illumina sometimes use different probe identifiers for the same probe sequences; in meta-analysis and Table 3, the label HT-12 was used if both HT-12 and Ref-8v2 had the same sequence.

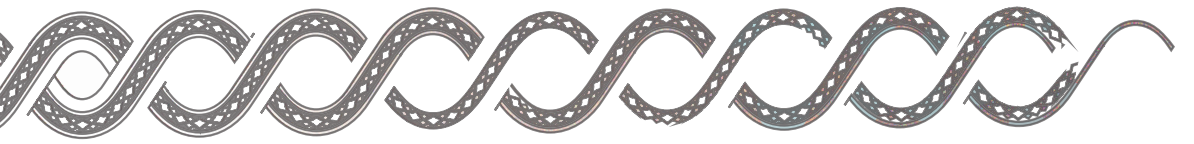
Re-mapping of probes. If probes mapped incorrectly or cross-hybridized to multiple genomic loci, it might be that an eQTL would be detected that would be deemed a trans-eQTL. To prevent this, we used a mapping approach versus a known reference that we developed for high-throughput short sequence RNAseq data[44]. We took the DNA sequence as synthesized for each cDNA probe and aligned it against a transcript masked gDNA genome combined with cDNA sequences. A more detailed description of probe re-mapping is provided in a Supplementary Note. Probes that did not map or that mapped to multiple different locations were removed.

Affymetrix transcriptional components applied to Illumina expression data. TC scores can be inferred in new (non-Affymetrix) datasets for every new individual sample. For the Illumina samples (used for the cis-eQTL mapping), only Illumina probes that could be mapped to any of our 22,106 Affymetrix probes were used (www.switchtoi.com/probemapping.ilmn). The *TCscore* t=n of sample *i* for the j^{th} TC is defined as: $TCscore_{ij} = \sum a_{ti} \times v_{tj}$, where v_{tj} *is t=1* defined as the t^{th} Affymetrix probe coefficient for the j^{th} TC; a_{ti} is the Illumina expression measurement for the t^{th} mapped probe for sample *i*. We inferred the Illumina TC scores for the top 1,000 TCs.

Removal of transcriptional component effects from Illumina expression data. Because our Illumina eQTL dataset (n = 1,469) is much less hetero- geneous than the Affymetrix dataset (n = 33,109), we expect that some TCs will

hardly vary. We therefore performed a PCA on the covariance matrix of the top 1,000 inferred TC scores for the Illumina dataset to effectively compress the TC data into a small set of 'aggregate TCs' (aTCs). As aTCs are orthogonal, we used linear regression to eliminate the effect of the top 50 aTCs. We correlated the TC-scores for each peripheral blood sample with probe expression levels. We then used the resulting residual gene expression data for subsequent cis-eQTL mapping.

cis-eQTL mapping. We used the residual gene expression data in a meta-analysis framework, as described[45,46]. In brief, analyses were confined to those probe-SNP pairs for which the distance from probe transcript mid-point to SNP genomic location was less than 500 kb. To prevent spurious associations due to outliers, a nonparametric Spearman's rank correlation analysis was performed. When a particular probe-SNP pair was present in both the HT12 and H8v2 datasets, an overall, joint P value was calculated using a weighted Z-method (square root of the dataset sample number). To correct for multiple testing, we controlled the false-discovery rate (FDR). The distribution of observed P values was used to calculate the FDR, by permuting expression phenotypes relative to genotypes 1,000 times within the HT12 and H8v2 dataset. Finally, we removed any probes from analysis which contained a known SNP (1000Genomes CEU SNP data, April 2009 release).

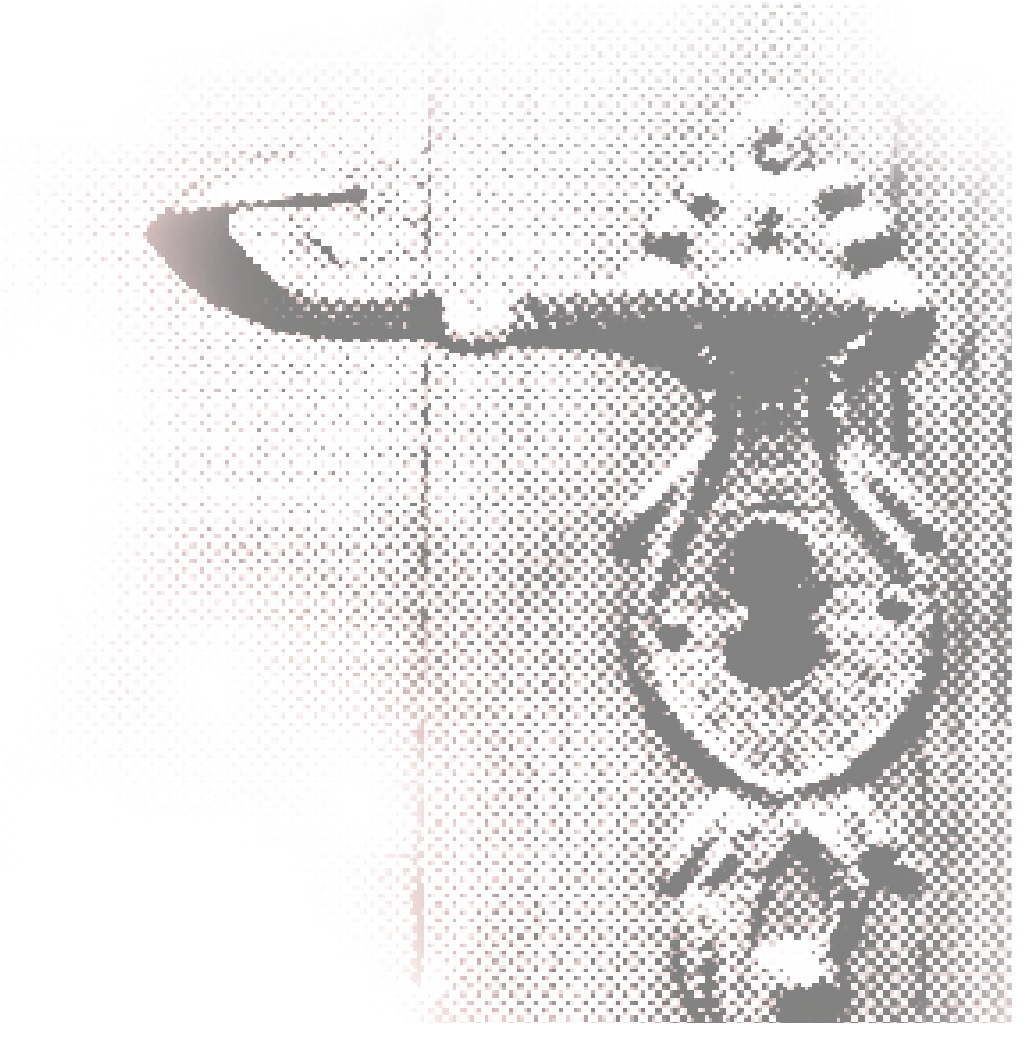


Chapter 5

A genetic perspective on coeliac disease

Trends Mol Med. 2010 Nov;16(11):537-50

Gosia Trynka, Cisca Wijmenga and David A van Heel



Coeliac disease is an inflammatory disorder of the small intestine with an autoimmune component and strong heritability. Genetic studies have confirmed strong association to HLA and identified 39 nonHLA risk genes, mostly immune-related. Over 50% of the disease-associated single nucleotide polymorphisms are correlated with gene expression. Most of the coeliac disease-associated regions are shared with other immune-related diseases, as well as with metabolic, haematological or neurological traits, or cancer. We review recent progress in the genetics of coeliac disease and describe the pathways these genes are in, the functional consequences of the associated markers on gene expression and the genes shared between coeliac disease and other traits.

Coeliac disease

Coeliac disease is an inflammatory disorder with an autoimmune component. The prevalence of coeliac disease in populations of white European origin is estimated at 1–3%[1]. The classic manifestation of coeliac disease includes diarrhoea, abdominal distension, failure to thrive and short stature[2]. Many patients present with extraintestinal manifestations, such as dermatitis herpetiformis, anaemia, neurological symptoms or osteoporosis[3,4]. This wide spectrum means coeliac disease is correctly diagnosed in only one in seven cases, hence only a minority of patients receives treatment. Coeliac disease is a life-long condition and the only known treatment is to completely exclude gluten from the diet. Approximately 5% of patients do not respond to a gluten-free diet and have a higher risk of developing enteropathy-associated T cell lymphoma[5]. Some but not all studies suggest that there is a higher mortality rate in patients who remain untreated[6–8].

The key molecule in mediating disease pathogenesis is gluten, a dietary protein found in wheat, barley and rye. Gluten peptides pass through the epithelial barrier of the intestine into the lamina propria, where they can be deamidated by the enzyme tissue transglutaminase. This enzymatic modification gives a negative charge to the gluten peptides and increases their immunogenicity[9].

In genetically predisposed individuals, gluten peptides trigger a cascade of innate and adaptive immune responses and lead to the destruction of the intestinal epithelium and mucosa as well as to lymphocytic infiltration in the proximal part of the small bowel[10]. In the lamina propria, gluten peptides are recognised by antigen-presenting cells (APCs) and presented by HLA class II DQ2 or DQ8 molecules[11,12]. Deamidated gluten peptides bind strongly to the specific HLA molecules[13]. The subsequent adaptive response is mediated by $CD4^+$ T cells that recognise the HLA–gluten complex and produce proinflammatory cytokines, mainly interferon (IFN)-γ[14]. This response leads directly to tissue remodelling and flattening of the intestinal mucosa[15]. Gluten peptides also trigger an innate immune response characterised by the elevated expression of interleukin (IL)-15 by intestinal enterocytes[16]. Consequently, patients with coeliac disease develop antibodies to gluten and autoantibodies to endogenous tissue transglutaminase, the dominant autoantigen[17]. Our knowledge of coeliac disease pathogenesis is expanding and we are beginning to understand certain parts of the disease process better, such as the pathogenicity of various gluten epitopes[18]. Although wheat products contain many different gluten pep- tides, only a small subset seems to be toxic to patients[18], which opens up new avenues for future therapies, such as vaccination. There are, however, many details of the disease process that remain unclear, such as the translocation of gluten peptides across the intestinal barrier and the process of flattening the intestinal mucosa.

Coeliac disease is a complex disorder in which genetic factors play an important role. The best characterised and most important genetic risk factor for coeliac disease is an HLA class II gene. Although HLA is necessary for coeliac disease to develop, it is not sufficient. Genome-wide association studies (GWAS) of coeliac disease have identified 39 nonHLA loci that also predispose to coeliac disease. The majority of these loci are not specific for coeliac disease but are also associated with other disorders.

Here, we summarise the latest achievements in coeliac disease genetics and describe the risk loci identified, the most likely disease genes in these loci, their pathways and their potential functional consequences on gene expression. We also discuss the genetic background shared between coeliac disease and other disorders.

Genetic insights into coeliac disease

The risk of HLA

HLA class II molecules are the major risk factors predis- posing individuals to coeliac disease and account for ~35% of the genetic risk[19]. Over

90% of patients with coeliac disease express the HLA-DQ2 heterodimer and the remainder express HLA-DQ8 molecules. HLA-DQ2 is encoded by the *HLA-DQA1*05* allele (a chain) and *HLA-DQB1*02* allele (b chain). The two alleles are often present in the *cis* conformation on the DR3 haplotype, which is also common to many other autoimmune disorders[20]. The small proportion of patients with coeliac disease who do not express *HLA-DQ2* molecules are *HLA-DQ8*-positive, where the a and b chains of the *HLA-DQ8* molecule are encoded by *HLA-DQA1*03* and *HLA-DQB1*0302*, respectively.

GWAS findings

Although the role of HLA molecules and the association to particular genotypes has been well established and explains their role in the disease pathogenesis, the frequency of coeliac risk HLA genotypes in the general population is ~30%, whereas only 1–3% actually develop the disease. As the heritability of coeliac disease is estimated to be ~80%[21] and HLA is estimated to contribute 35% of the genetic risk[19,22], there must be more genetic risk factors involved in coeliac disease susceptibility. This is corroborated by twin studies that show a large discrepancy in the concordance rate of coeliac disease in monozygotic twins compared to HLA-identical dizygotic twins[21].

A breakthrough in finding non-HLA coeliac risk genes was achieved with the advent of GWAS (Box 1) and their application to complex diseases. Thus far, the two GWAS that have been conducted on coeliac disease have reported definite associations to 26 nonHLA loci and have suggested another 13 loci as predisposing to coeliac disease (Table 1)[19,22–26].

Understanding the functional role of coeliac risk variants: coeliac e-QTL mapping

Although GWAS have identified common variants underlying susceptibility for common complex traits, the interpretation of the findings requires further study to identify causal variants and genes. GWAS utilise a tag single nucleotide polymorphism (SNP) approach, meaning that the genetic markers that are tested for association to a trait or disease are selected to capture the majority of the genetic variation. This is possible because of the extensive allelic association between SNPs in close proximity to each other in the human genome [i.e. linkage disequilibrium (LD)] (Box 1). Consequently, only a subset of SNPs that are informative enough to capture the majority of the variation of the untyped SNPs needs to be genotyped. The obvious drawback of this approach is that when an association is observed, it is unclear what the underlying causal variant is. Moreover, if the association maps within a region with multiple genes in strong LD with each other, it is difficult to pinpoint the true risk gene. Sequence analysis of entire associated regions can help identify causal variants[27], although the nature of causal variants in complex diseases remains to be established.

One way to suggest possible causal genes is by assessing the correlation between genotypes of the associated SNPs and the levels of expression of genes in this same region (e- QTL mapping, Box 2). Such a functional genomics approach has shown that 53% of the coeliac disease-associated risk regions (i.e. 20 of 38 tested non-HLA loci) contain SNPs with a significant cis e-QTL effect in peripheral blood mononuclear cells (PBMCs). To date, many eQTL studies have been performed using PBMCs because of the ease of collecting them. Consequently, the expression data comes from a limited set of cell types, which might not always be relevant to the disease under study. It cannot be excluded that genes with expression specific to intestinal epithelial cells, for example, are also relevant for coeliac disease pathogenesis and these will be missed using PBMCs. It also is important to acknowledge that e-QTL mapping does not prove causality, but identifying a SNP with a functional effect on a nearby gene can help to prioritise genes for functional follow-up studies, for example in regions where there are multiple genes in tight LD. One such example is 2q12.1, where several genes encoding IL receptors are present but an e-QTL effect is only detected for *IL18RAP*. Interestingly, for some regions, an associated

SNP has both a negative and a positive e-QTL effect: for example at 1p36.32, the rs3748816*A allele corresponds to down-regulation of *PLCH* and *TNFRSF14* but upregulation of *MMEL1* and *C1orf93*. For 71% of coeliac e-QTL SNPs, the associated coeliac risk allele corresponds to a downregulation of gene expression (Table 2), suggesting that most coeliac risk alleles lead to a reduction in gene expression. However, although e-QTL mapping helps to indicate possible causal genes, the results need to be interpreted with caution. There is a greater overrepresentation of the e-QTL SNPs among coeliac-associated loci than would be expected by chance (22 of 44 tested SNPs compared to 7.8 expected eQTL SNPs by chance), but some of the detected effects are still likely to result from chance thus some of them are probably not related to the disease aetiology.

Coeliac disease pathways

The 39 nonHLA coeliac loci together encompass 115 different genes (based on LD blocks[19]). At least 28 of the non-HLA coeliac loci harbour immune-related genes, pointing to an altered immune system response underlying coeliac disease. Based on our knowledge of coeliac disease biology, the most plausible immune-related genes from these 28 regions can be grouped into several immune-related pathways. Many of these genes have broad functions and regulate several pathways. Below we describe how GWAS findings have enhanced our knowledge of the key coeliac disease pathways identified so far.

Adaptive immune response

Coeliac disease has long been recognised as a T cell disease, with a strong adaptive immune response. This view was mainly based on the genetic association with HLA, which was established more than 30 years ago. Based on the GWAS results, specific genes can now be identified that provide a much broader picture, including the triggering and activation of T cells and a role for the thymus. The latter observation is intriguing as it provides a possible link to the lack of oral tolerance to dietary gluten seen in patients with coeliac disease.

Immune cell signalling

Immune cell signalling is a dynamic process that involves immune cells interacting with other cells, presenting antigens as well as responding to secreted factors such as cytokines. Among the coeliac disease-associated genes are those that modulate immune cell signalling at different levels (Figure 2a). Genes such as *CTLA4*, *SH2B3*, *PTPN2* or *CD80* encode factors that negatively regulate T cell responses, whereas others are involved in the T cell mediated cell apoptosis (such as *FASLG*). Many of the coeliac loci contain genes encoding factors affecting chemokines (*CCR1, CCR3, CCR4, CCR5, CCR9, CXCR1, CXCR6*), which in turn regulate immune responses by directing cellular migration towards sites of inflammation or to wards lymphoid tissues. By blocking the Jak (Janus kinase)/STAT (signal transducers and activators of transcription) signalling pathway, *SOCS1* (suppressor of cytokine signalling-1), encoded by another coeliac gene, negatively regulates cytokine signalling[28]. The *SH2B3* locus, which shows signs of positive selection, encodes a factor likely to be involved in inhibiting cytokine responses. This gene was recently implicated in NOD2 (nucleotide-binding oligomerisation domain containing 2) signalling as carriers of the *SH2B3*-rs3184504*A risk allele show an increased proinflammatory immune response upon lipopolysaccharide and muramyl dipeptide stimulation. This increased response also suggests that *SH2B3*-rs3184504*A carriers are protected against bacterial infections[29], which might have driven evolutionary selection. Another coeliac gene, *ICOSLG*, encodes a ligand for the ICOS receptor that blocks the interaction between T cells and APCs and therefore inhibits the antibody-specific T cell response. This ligand is expressed on activated T cells and regulates expression of IL-21 in the development of Th17 cells[30]. ICOSLG costimulation induces production of IL-10 or IFN-g[31]. CD247 encodes the zeta chain, which is part of the T cell receptor CD3 complex and plays an important role connecting antigen recognition with downstream signal transduction[32].

T cell maturation and differentiation

Box 1. GWAS and e-QTL

GWAS

GWAS genotype hundreds of thousands to millions of SNPs across a large set of DNA samples, with the aim of identifying SNPs with significantly different frequencies between cases (carriers of the trait of interest) and controls. A p-value of $5x10^{-8}$ is generally the level needed for statistical significance. Genotyped SNPs are selected to capture the majority of common variation in the human genome. This is possible because of the LD patterns seen in the human genome, which means that SNPs in proximity generally provide information about each other. Thus, it is not necessary to genotype all of the SNPs; a subset are genotyped and the remainder of the genetic information is inferred by imputation methods. These methods compare the patterns of genotyped markers (from GWAS data) with those present in a reference panel (e.g. HapMap) and the missing genotypes are predicted based on the observed similarities[60].

***cis* e-QTL mapping for disease-associated SNPs**

Associated SNPs that reach a genome-wide significance level ($p<5x10^{-8}$) often map within regions of tight LD, where several genes can be present. Correlation between the genotypes of the associated SNP and the expression levels of the genes in its proximity can help to pinpoint the best candidate gene. In this type of analysis, there are three possible scenarios: positive, negative or no correlation. Assuming B is the risk allele, the most associated SNP can correlate positively with expression of gene X, have no effect on the expression level of gene Y, or result in low expression of gene Z. A single SNP can affect the expression of one or more genes in the locus; the SNP can affect the expression of multiple genes in the same direction or in the opposite direction. Conversely, the expression of a single gene can be affected by multiple SNPs and can be negatively regulated by one SNP but positively by another. Although the correlation between genotype and gene expression shows a functional relationship between a SNP and a gene, it is not proof of causality between a gene and the disease under study (Figure I).

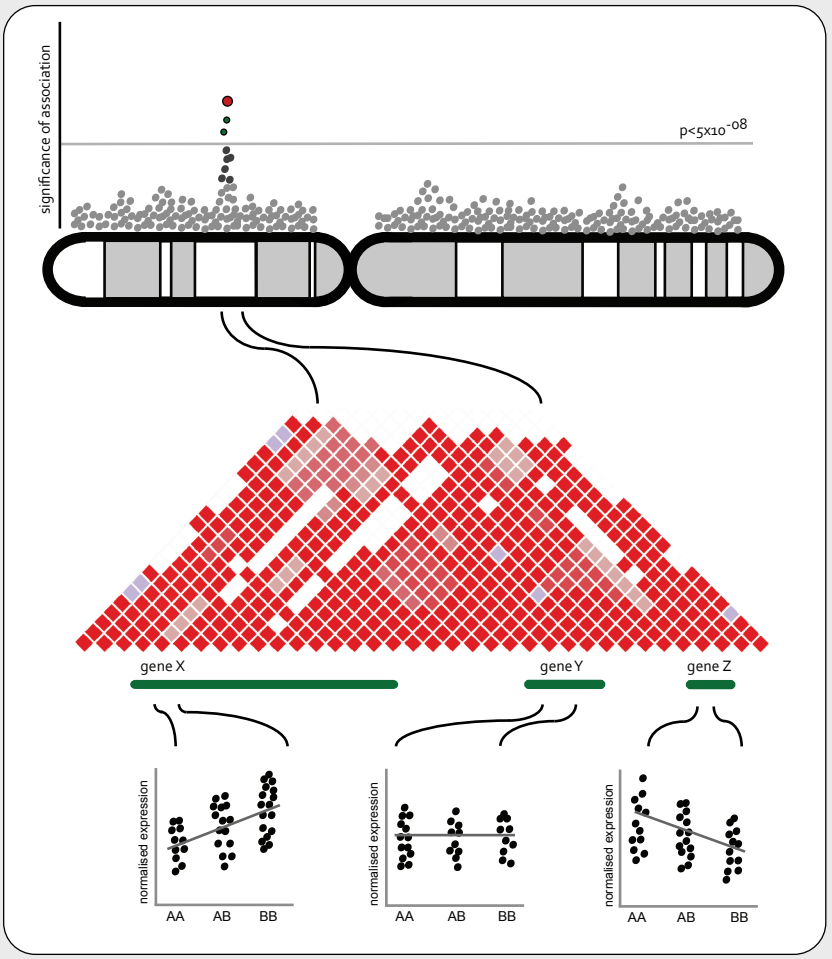


Table 1. Coeliac disease susceptibility genes: function, pathways and associations to other traits

Disease/trait	Region	Genes in the region	Function of the best candidate gene	Pathway	Ref.
Coeliac disease Chronic kidney disease Diastolic blood pressure Esophageal cancer Hematocrit Hematological and biochemical traits Hemoglobin Plasma eosinophil count Systolic blood pressure Type 1 diabetes	12q24.12*	*CUX2, FAM109A, ATXN2, TMEM116, SH2B3, RPL6, ERP29, TRAFD1, ACAD10, PTPN11, MAPKAPK5, BRAP, ALDH2*	SH2B3: regulation of T-cell receptor, growth factor and cytokine receptor-mediated signalling	Immune cell signalling	17 58 59,60 61 62 63 62 64 59 65
Coeliac disease Inflammatory bowel disease (early onset) Multiple sclerosis	10q22.3	*ZMIZ1*	ZMIZ1: member of the family of protein inhibitor of activated STAT; regulates activity of several transcription factors, modulates transforming growth factor-β signalling	Immune cell signalling	17 66 67
Coeliac disease Crohn's disease Type 1 diabetes	18p11.21	*PTPN2*	T cell protein tyrosine phosphatase-non receptor 2, suppresses IL6 induced STAT3 tyrosine phosphorylation and activation	Immune cell signalling	17 68-70 65,71, 72
Coeliac disease Rheumatoid arthritis Type 1 diabetes	2q33.2	*CTLA4, ICOS*	CTLA4: negative regulator of T cell responses. ICOS: inducible T-cell co-stimulator precursor, belongs to CTLA-4 cell-surface receptor family	Immune cell signalling	17 73 65,72
Coeliac disease Type 1 diabetes	14q24.1	*ZFP36L1*	ZFP36L1: putative nuclear transcription factor may function in regulating the response to growth factors	Immune cell signalling	17 65
Coeliac disease Type 1 diabetes	16p13.13	*TNP2, PRM3, PRM2, PRM1, **SOCS1**, CLEC16A*	SOCS1: suppressor of cytokine signalling 1; its expression can be induced by IL2 and IFNγ. Takes part in a negative feedback loop to attenuate	Immune cell signalling	17 72
Coeliac disease Systemic sclerosis	1q24.2	***CD247***	Part of the CD3 receptor complex where it plays an important role in coupling antigen recognition to several intracellular signal-transduction pathways. Low expression of CD247 results in the impaired immune response	Immune cell signalling	17 74
Coeliac disease Crohn's disease	21q22.3	*ICOSLG, RRP1///AP001053.1*	ICOSLG: inducible T cell co-stimulator ligand; regulates T cell activation, up-regulates IL10 expression. ICOS receptor is a member of CTLA4/CD28 family. RRP1: involved in the late stages of nucleologenesis at the end of mitosis	Immune cell signalling	17 69
Coeliac disease	1q31.2	*RGS1*	Regulates chemokine receptor signalling and is involved in B-cell activation and proliferation. Specifically expressed in intestinal intra-epithelial lymphocytes	Immune cell signalling	17
Coeliac disease	3q13.33	*CDGAP, TMEM39A, KTELC1, **CD80***	CD80: can bind to CTLA-4 which results in the decreased response of the T-cells. CD80 is expressed on activated B cells and monocytes	Immune cell signalling	17
Coeliac disease	6q25.3	*RSPH3, TAGAP*	TAGAP: a RhoGTPase-activating protein, expressed in activated T-cells and important for modulating cytoskeleton changes	Immune cell signalling	17
Coeliac disease Crohn's disease	1q24.3	***FASLG**, TNFSF18, TNFSF4*	FASLG: cytokine that binds to FAS. Involved in cytotoxic T-cell mediated apoptosis and in T-cell development. TNFSF18: modulates T cell survival in the peripheral tissues. TNFSF4: involved in T cell antigen-	Immune cell signalling	17 68,69

Celiac disease Systemic lupus erythematosus	11q24.3	*ETS1*	A key player during CD8 lineage differentiation in the thymus via direct promotion of RUNX3 expression	T-cell maturation and differentiation	17 75,76
Coeliac disease Rheumatoid arthritis	1p36.32	*PANK4, MMEL1, PLCH2, HES5, TNFRSF14*	MMEL1: member of the membrane metallo-endopeptidase (MME) family. TFRSF14; tumor necrosis factor receptor superfamily member 14, part of herpes simplex virus (HSV) entry mechanism, promotes apoptosis of double-positive thymocytes	T-cell maturation and differentiation	17 77
Coeliac disease	1p36.11	*RUNX3*	Promotes differentation of CD4+ cells into Th1 cells. Master regulator of CD8+ T lymphocyte developement in the thymus	T-cell maturation and differentiation	17
Coeliac disease	6q22.33	*PTPRK, THEMIS*	THEMIS: plays a regulatory role in both positive and negative T-cell selection during late thymocyte development. The protein functions through T-cell antigen receptor signalling, and is necessary for lineage commitment and maturation of T-cells.	T-cell maturation and differentiation	17
Coeliac disease Plasma eosinophil count	2q12.1*	***IL18RAP**, IL1RL2, IL18R1, SLC9A4, IL1RL1*	IL18RAP: subunit of IL18 receptor. Cytokine of broad function including stimulation of interferon gamma production by T-cells. Role in activation of NF-κB and JNK in response to IL18	T-cell maturation and differentiation	17 64
Coeliac disease Multiple sclerosis	3q25.33*	*SCHIP1, IL12A*	IL12A: required for the T-cell-independent induction of IFNγ; involved in differentiation of Th1 and Th2 cells	T-cell maturation and differentiation	17 67
Coeliac disease Type 1 diabetes	4q27	*IL2, IL21, ADAD1, KIAA1109*	IL2: T-cell growth factor; activation and proliferation of NK cells, monocytes, macrophages, differentiation of B cells. IL21: Amplification of Th17 response	T-cell maturation and differentiation	17 65,70
Celiac disease Hematological and biochemical traits Systemic lupus erythematosus	22q11.21	*CCDC116, UBE2L3, HIC2, LOC150223*	UBE2L3: ubiquitin-conjugating enzyme E2L3, participates in the ubiquitination of p53, c-Fos, and the NF-κB precursor p105 in vitro	Innate: NF-kB	17 63 76
Coeliac disease Rheumatoid arthritis Ulcerative colitis	2p16.1	*REL, PUS10, PEX13, AHSA2*	REL: a component of the NF-kB transcription complex that plays a critical role in promoting immune and inflammatory responses through the production of pro-inflammatory cytokines AHA1: activator of heat shock 90kDa protein	Innate: NF-kB	17 73 78
Ankylosing spondylitis Coeliac disease	2q31.3	*UBE2E3, ITGA4*	UBE2E3: by conjugating with SUMO-1 negatively regulates NF-kB activity. ITGA4: encodes an alpha subunit of the integrin, acts as adhesive and signalling receptor for leukocytes	Innate: NF-kB	79 17
Coeliac disease Rheumatoid arthritis	6q23.3	*OLIG3, TNFAIP3*	OLIG3: neuronal developement. TNFAIP3: zinc finger protein, inhibits NF-κB activation as well as TNF-mediated apoptosis; limits inflammation by terminating TNF-induced NF-κB responses.	Innate: NF-kB	17 77,80
Celiac disease	1p36.23	*TNFRSF9, PARK7, ERRFI1*	TNFRSF9: member of the TNF receptor superfamily. T lymphocyte co-stimulation to promote Th1 response. Activates NF-κB. PARK7: protects neurons against oxidative stress and cell death.	Innate: NF-kB	17
Brain structure Coeliac disease	3p21.31	*CCR5, CCR3, LTF, CCRL2, CCR1, CCR9, CXCR6, XCR1, CCRL2,* [g]	Chemokine receptor cluster, chemokines are involved in recruitment of effector immune cells to the site of inflammation	Chemokines	81 17

Coeliac disease	3p22.3	*TRIM71, CCR4, GLB1*	CCR4: chemokine receptor, one of the ligands is RANTES which recruits leukocytes into inflammatory sites.	Chemokines	17
Black vs. blond hair color Black vs. red hair color Coeliac disease Chronic lymphocytic leukemia Freckles Tanning	6p25.3	***IRF4***	IRF4: transcriptional activator, binds to the interferon-stimulated response element (ISRE) and plays a role in ISRE-targeted signal transduction mechanisms. Activated in TLR7 and TLR9 signalling pathways, expression of TRF4 is specific to T-, B- cells and macrophages. Also crucial for IL-21 mediated activation and stabilization of Th-17 cells.	Innate: response to virus infection	82 82 17 83 84 85
Coeliac disease Type 1 diabetes	6q15	*BACH2*	Transcriptional repressor in B-cells, a key regulator of antibody response. Regulator of nucleic acid-triggered antiviral responses	Innate: response to virus infection	17 65,72, 86
Coeliac disease	Xp22.2	*TLR8, TMSL3, TLR7, TMSB4X*	TLR7/TLR8: play fundamental role in recognition of single stranded viral RNA and activation of innate immunity.	Innate: response to virus infection	17
Coeliac disease Parkinson's disease	17q21.31	*MAPT, KIAA1267, LRRC37A, ARL17B, NSF, WNT3*	NSF: participates in vesicle-mediated transport. WNT: WNT family encodes secreted signalling proteins. Implicated in oncogenesis and in several developmental processes	Other	17 87
Coeliac disease Serum metabolites	2p14	*PLEK*	Pleckstrin: platelet and leukocyte C kinase substrate, associates with membranes in human platelets and may affect membrane structure.	Other	17 88
Coeliac disease Vitiligo	3q28	*LPP*	LPP: localizes to the cell periphery in focal adhesions and may be involved in cell-cell adhesion and cell motility.	Other	17 89
Coeliac disease	1p31.3	*NFIA*	Plays a role in regulation of human granulopoiesis.	Other	17
Coeliac disease	3p14.1	*FRMD4B*	FERM domain-containing protein 4B. May provide a link between membrane to cytoskeleton and be involved in signal transduction	Other	17
Coeliac disease	7p14.1	*ELMO1*	Promotes phagocytosis and affects cell shape changes	Other	17
Coeliac disease Telomere length	3q26.2	*ARPM1, LRRC34, LRRC31, MYNN, LOC344657*	The function of these genes is unknown	Unknown	17 90
Coeliac disease Crohn's disease Ulcerative colitis	1q32.1	*Intergenic*	Unknown	Unknown	17 69 69,78
Coeliac disease	13q14.2	*Intergenic*	Unknown	Unknown	17
Coeliac disease	8q24.21	*Intergenic*	Unknown	Unknown	17

Underlined genes have a detected e-QTL effect; genes in green are downregulated and in red are upregulated in association with the risk allele. Independent SNPs that have an e-QTL effect on the same gene but with opposite effects are in blue. Genes identified from GRAIL analysis [19] are in bold font. Regions with signatures of positive selection are marked with asterisks.

Individuals with immune-related diseases have a high number of T cells as well as an imbalance in T cell subsets compared to normal individuals. From the perspective of immune-related diseases, there are three subtypes of CD4 cells of particular interest: Th1, Th17 and regulatory T cells (Tregs). T helper cells coordinate the adaptive immune response against pathogens by producing specific cytokines and thereby mediating different types of tissue inflammation. Both the IL-12 and IL-18 pathways induce IFN-γ secretion by Th1 cells; IFN-γ is the key cytokine involved in the mucosal inflammation in coeliac disease. The GWAS results suggest an important role for Th1 cells as they identified two regions associated with coeliac disease that harbour genes in both the IL-12 and IL-18 path- ways: the 2q12.1 region containing the *IL18RAP* and *IL18R1* genes, and the 3q25.33 region containing *IL12A*, which encodes a subunit of the heterodimeric IL-12 cytokine. Th17 cells might be potent inducers of autoimmunity and tissue inflammation[33]. The GWAS results suggest a role for either

Th17 or Treg cells as the chromosome 4q27 locus contains both the *IL2* and *IL21* genes. IL-21 stimulates expansion of Th17 cells[30] and regulates inflammation and Th17 responses in the gut[34]. IL-2 is a general T cell growth factor and activates Treg cells, a subset of T cells that can suppress the responses of effector T cells and other immune cells. Decreased activity or altered functional properties of Treg cells result in severe autoimmunity[35].

Thymus

The induction of central tolerance as well as CD4 and CD8 lineage commitment from thymocytes takes place in the thymus[36]. The GWAS results implicate the thymus in the pathogenesis of coeliac disease (Figure 2). For example, *THEMIS* encodes a protein involved in T cell maturation and in regulating lineage commitment of thymocytes into CD4 or CD8[37] cells, whereas *ETS1* and *RUNX3* genes both encode regulators of CD8 T lymphocyte development[38]. *TNFRSF14* encodes a protein that promotes apoptosis of double-positive thymocytes[39].

Innate immune response

Box 2. Pathway analysis tools

Recent advances in technology have provided researchers with high-throughput methods, such as GWAS, genome-wide gene expression analysis and whole-exome sequencing, providing a great capacity to study biological processes involved in disease pathogenesis. However, such analyses often results in long lists of potentially interesting genes. Applying pathway analysis tools such as DAVID to a list of candidate genes can help in interpreting the results. Such tools map the genes of interest to the associated, annotated biological pathways and reveal the most enriched annotations. Pathway analysis tools utilise different biological annotation resources (e.g. KEGG, BioCarta and gene ontology) all of which aim to integrate gene information into biological pathways. DAVID, the Database for Annotation, Visualisation and Integrated Discovery (http://david.abcc.ncifcrf.gov/), provides a comprehensive set of functional annotation tools to understand biological meaning behind large lists of genes[61]. The application of any pathway analysis tool has limitations. The major drawback is that analysis of a random selection of genes emerging from GWAS findings will always result in significantly overrepresented pathways. Therefore, any results should be treated with caution and as an indication of biologically relevant pathways for the disease rather than as proof of causality[55].

The role of the innate immune response in coeliac disease has been unclear; however, the release of IL-15 by intestinal epithelial cells in direct response to one of the gluten peptides has been described[40]. In addition, viral infections have been implicated as triggering factors in coeliac disease, although this is mainly anecdotal. The GWAS findings point to a causal role for the innate immune system as multiple genes are involved in NF (nuclear factor)-κB and Toll-like receptor (TLR) signalling.

Innate response to infection

GWAS point to the importance of the innate response to exogenous factors, such as infection (Figure 2). Genes such as *TLR7/TLR8, BACH2* and *IRF4* indicate the involvement of an immune response triggered by viral infection. TLRs play a key role in activating the innate immune response in response to molecules from pathogens: TLR7 and TLR8 recognise viral RNA, whereas IRF4 is a transcriptional activator in the TLR7 signalling pathway. IRF4 is specific to macrophages, T and B cells[41]. BACH2 is the key regulator of antibody response in nucleic acid-triggered antiviral response. Interestingly, *BACH2* is also associated with type 1 diabetes, which makes it tempting to speculate that viral infection could be a trigger for both of these diseases; such a mechanism triggers a multiple sclerosis phenotype in a mouse model[42,43]. A high frequency of rotavirus infections might increase the risk of developing coeliac disease in susceptible individuals[44].

Innate NF-kB signalling

NF-κB is a transcriptional complex that regulates adaptive and innate immune responses, inflammation through cytokine production and cell death[45]. Genomic regions associated with coeliac disease contain genes encoding factors that regulate NF-κB signalling. For example, *REL* encodes a component of the NF-κB complex, whereas the *TNFAIP3* gene product A20 terminates NF-κB activity. UBE2E3, through conjugation with SUMO-1, negatively regulates NF-κB and UBE2L3 ubiquitinates an NF-kB precursor in vitro, possibly affecting protein stability.

TNFRSF9 encodes a member of the tumour necrosis factor (TNF) receptor superfamily that can activate NF-κB. The association of these genes strongly suggests that NF-κB signalling plays a role in the development of coeliac disease pathology.

Other coeliac disease pathways

For a subset of genes, little is known about the function of their products, and for others association signals map to intergenic regions (Table 1). Because these genes and loci do not immediately fit into our current understanding of coeliac disease pathogenesis, they present new leads for studying the disease aetiology.

Individual risk profiling for coeliac disease

Similar to the majority of complex traits evaluated so far by GWAS, the non-HLA genes show a modest individual risk and all 39 loci together explain only ~5% of the risk for coeliac disease[19]. Although 5% might seem to account for only a small proportion of the total genetic risk, it can be used to assess an individual's risk for coeliac disease. The first step in determining individual risk is to assess their HLA status because the absence of *HLA-DQ2/DQ8* is a strong predictor of not developing coeliac disease (close to zero risk). One or two copies of *HLA-DQ2* give an a priori intermediate or high risk for disease. However, the non-HLA risk alleles determine the absolute risk, which can increase to >30%[46,47]. Before genetic profiling can be implemented in clinical practice, the specificity and sensitivity of the genotyping profiles must be improved and the positive predictive value needs to be determined using prospective and longitudinal cohort studies. Nevertheless, if in the future, we can identify individuals at risk, this might allow for early intervention[48], especially in families that already have patients with coeliac

Table 2. eQTL effects of the coeliac disease associated SNPs

Chromosome	SNP	Risk allele	eQTL gene	Expression change
1	rs3748816	A	C1orf93	Up
1	rs3748816	A	PLCH	Down
1	rs3748816	A	MMEL1	Up
1	rs3748816	A	TNFRSF14	Down
1	rs12727642	A	PARK7	Up
1	rs864537	A	CD247	Down
1	rs864537	A	CD247	Down
1	rs296547	G	Probe-rs296547	Down
2	rs842647	A	AHSA2	Down
2	rs13003464	G	AHSA2	Down
2	rs3816281	C	PLEK	Down
2	rs917997	A	IL18RAP	Down
2	rs13010713	G	UBE2E3	Up
3	rs13098911	A	CCR3	Down
3	rs13098911	A	CCR3-probe2190671	Down
3	rs13098911	A	Probe-6550333	Down
3	rs13098911	A	CCR3	Down
3	rs6441961	A	CCR3	Up
3	rs11922594	G	KTELC1	Down
3	rs11922594	G	KTELC1-probe6550288	Down
6	rs10806425	A	BACH2	Down
6	rs1738074	A	probe-5890739	Down
6	rs1738074	A	TAGAP	Down
6	rs1738074	A	TAGAP-probe5360364	Down
7	rs6974491	A	ELMO1	Down
10	rs1250552	A	ZMIZ1	Down
12	rs653178	G	SH2B3///ATXN2	Up
12	rs653178	G	ALDH2	Down
12	rs653178	G	TMEM116	Down
12	rs653178	G	TMEM116-probe2070736	Down
16	rs12928822	G	Probe-4540072	Down
17	rs2074404	A	ARL17///LRRC37A///LRRC37A2	Down
17	rs2074404	A	NSF-probe5260138	Down
17	rs2074404	A	WNT3	Down
17	rs2074404	A	NSF	Down
17	rs2074404	A	Probe-4880037	Up
21	rs4819388	G	RRP1///AP001053.1	Down
22	rs2298428	A	Probe-1230242	Up
X	rs5979785	A	TLR8	Up
X	rs5979785	A	TLR8-probe3390612	Up

A total of 22 coeliac disease-associated SNPs with e-QTL effects[19]. In total, 23 e-QTL events result in downregulation of gene expression and 9 in upregulation. SNPs affecting expression of several probes located within the same gene were counted only once.

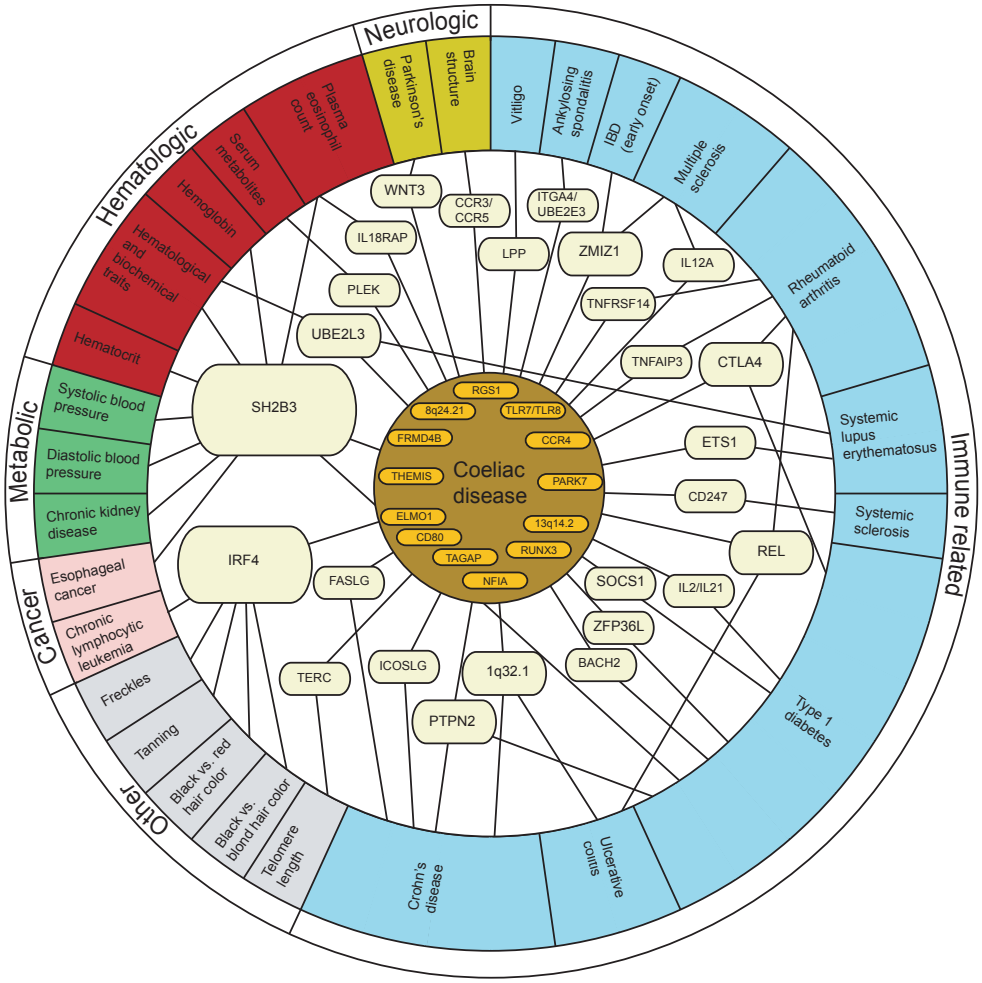


Figure 1. Genetic sharing of coeliac disease. The disease associations that fall to the same LD blocks as the 39 nonHLA coeliac disease loci were extracted from 'A Catalogue of Published Genome-Wide Association Studies'. A total of 26 loci were shared by one or more traits (presented in the modules in the circle). The size of the 'gene module' corresponds to the number of connections with traits. *SH2B3* is the most shared locus; it is associated with most of the haematological traits, metabolic traits and esophageal cancer. The size of the 'disease module' corresponds to the genetic sharing with coeliac disease, the bigger the module the more genes are in common. Immune-related diseases are overrepresented; in particular, type 1 diabetes, rheumatoid arthritis and Crohn's disease associate with four or more genes that are shared with coeliac disease. In total, 13 genes (in the central circle, orange) are specific for coeliac disease.

disease. Alternatively, risk profiling might assist monitoring by identifying individuals who should be repeatedly screened for coeliac disease-associated antibodies and, if necessary, have a biopsy taken.

Genetic sharing of coeliac disease with other traits

Many coeliac disease risk loci are shared with other immune-related diseases[25,49,50], which supports observations that autoimmune diseases co-occur in families and individuals[51]. Coeliac disease often co-occurs with type 1 diabetes, rheumatoid arthritis, ulcerative colitis or Crohn's disease[51–53].

A shared genetic background among these diseases points to common biological pathways underlying their aetiology. To obtain an unbiased picture of the shared genetics, we used 'A Catalogue of Published Genome-Wide Association Studies' (data freeze 19 May,

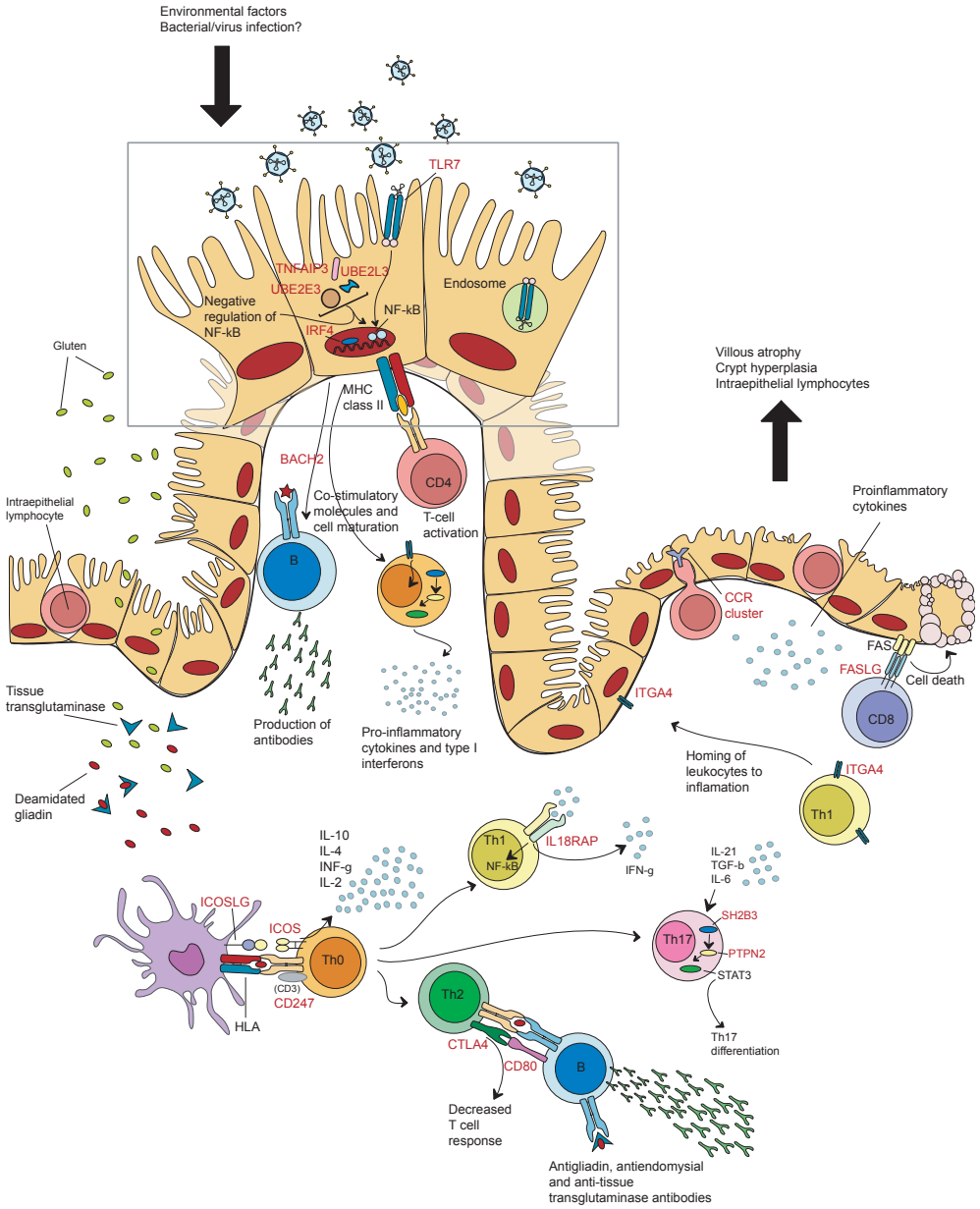


Figure 2. Possible mechanisms underlying coeliac disease pathogenesis. (a) Gluten peptides pass from the intestinal lumen through the impaired epithelial barrier into the lamina propria, where they undergo deamidation and are recognised by APCs. Deamidated gluten peptides bind efficiently to the HLA-DQ2 or DQ8 molecules on the surface of the APCs and are presented to T cells. This initiates the cascade of innate and adaptive immune responses and leads to inflammation of the small intestine, villous atrophy, crypt hyperplasia and infiltration of intraepithelial lymphocytes. Coeliac disease-associated genes might alter the immune response at different levels (associated candidate genes are in red). The GWAS findings indicate the causal role for the innate immune system. Multiple associated genes are involved in regulating the activity of NF-kB. Association of disease with loci harbouring genes involved in immune response to infection, *TLR7/TLR8*, *BACH2* and *IRF4*, indicate viral infection is a potential trigger for coeliac disease. (b) Induction of central tolerance together with $CD4$/CD8 lineage commitment from thymocytes occurs in the thymus. Genes associated with coeliac disease indicate that the thymus might be involved in the pathology. Associated genes regulate the fate of thymocytes by directing them towards negative or positive selection (e.g. *TNFRSF14* promotes apoptosis of double-positive thymocytes). Genes such as *ETS1* or *RUNX* are essential for the lymphocyte lineage commitment.

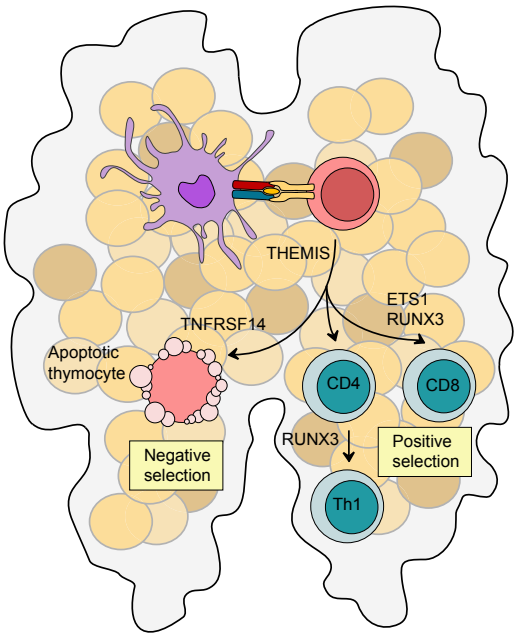


Figure 2. Continued

2010)[54] to extract all the reported trait and disease associations that map to the same LD blocks as the 39 coeliac disease-associated loci (for LD blocks definition see[19]). Overall, 26 loci were shared by one or more traits. The largest genetic sharing occurred between coeliac disease and type 1 diabetes (seven regions), followed by rheumatoid arthritis and Crohn's disease (four regions) (Figure 1). Interestingly, the *SH2B3* locus was associated with nine traits (apart from coeliac disease), most of which are haematologically related (Figure 1). The *IRF4* region was associated with five traits, most of which are unlikely to be relevant to coeliac disease (e.g. hair colour or freckles). Genes that are shared by three traits include *ZMIZ1, PTPN2*, an intergenic region on 1q32.1, *UBE2L3, REL* and *CTLA4*; all but the 1q32.1 locus are immune-related genes. Apart from its classic presentation, coeliac disease can also manifest as a skin phenotype (dermatitis herpetiformis), a neurology-related disorder (e.g. ataxia) and a disorder progressing into cancer (enteropathy-associated T cell lymphoma). Loci such as *CCR3/CCR5* or *WNT3* are shared with neurological traits, whereas *IRF4* and *SH2B3* are shared with cancers, suggesting that complications associated with coeliac disease could be partly explained by the shared genetics.

At this time, 13 loci remain private for coeliac disease. These results might be temporary as not all the disease-associated loci have been identified and different studies might have variable power to identify these loci because of sample size limitations or the effect sizes of one trait compared to another.

Pathway classification

We used our GWAS results to further our understanding of known pathways involved in coeliac disease by selecting relevant, immune-related genes from the different associ- ated loci. As such an approach is highly biased, we also applied DAVID, a pathway analysis tool (Box 2). We acknowledge that these tools are especially biased towards detecting the well-defined pathways[55] when analysing a random selection of SNPs. Although this analysis can indicate the biological processes involved in the disease aetiology, the results should be interpreted with caution.

Cellular processes involved in coeliac disease pathology

We applied DAVID analysis to the set of 115 genes selected from the 39 nonHLA loci. Of the 115 genes, 27 mapped to the KEGG (Kyoto encyclopaedia of genes and genomes) pathways and 17 genes lie on the cytokine–cytokine receptor interaction pathway (*CCR1, CCR3, CCR9, CCR4, CCR5, XCR1, CXCR6*) and on the TNF family subpathways (*FASLG, TNFSF4, TNFSF18, TNFRSF14, TNFRSF9*). Chemokine signalling involves the activation of the Jak/STAT, NF-κB and MAPK (mitogen-activated protein kinase) signalling pathways. The activation of NF-κB results in cytokine production, cell growth, differentia- tion, migration and apoptosis. The Jak/STAT signalling pathway is the main signal transduction mechanism down-stream of many cytokines and growth factors. Upon activation via Jak, STATs modulate the expression of the target genes. *SOCS1, IL12A, IL21, IL2* and *PTPN11* are implicated in this pathway.

CD80, *CCR9*, *ICOSLG*, *ICOS*, *ITGA4* and *IL-2* mapped within the intestinal immune network for immunoglobulin (Ig)A production, the second most significant pathway emerging from this analysis. IgAs are a class of noninflammatory immunoglobulins that is largely produced in the intestine and serves as the first-line defence against microorganisms. Multiple cytokines such as IL-10, IL-4, IL-5, IL- 6 and TGF (transforming growth factor)-β promote IgA production. Secreted IgAs capture dietary antigens and microorganisms in mucus and neutralise them. This result validates observations that selective IgA deficiency is often associated with coeliac disease[56] and now points to its genetic basis.

Other gene-enriched pathways identified by this approach include allograft rejection, type 1 diabetes mellitus, autoimmune thyroid disease, cell adhesion molecules, graft-versus-host disease and TLR signalling.

Shared pathways driving immune-related and inflammatory diseases

To investigate whether the pathways that drive coeliac disease are common to other immune-related diseases, we used 'A Catalogue of Published Genome-Wide Association Studies' (data freeze 19 May, 2010), focusing on the ten immune-related traits that share part of their genetics with coeliac disease and extracting all of the associated genes for pathway analysis that have been reported (using DAVID annotation).

Excluding HLA, there were 203 unique genes. Similar to coeliac disease, the cytokine-cytokine interaction pathway is the most enriched and includes almost 10% of the genes tested, followed by the Jak/STAT signalling pathway. Interestingly, two pathways enriched for coeliac disease genes, the chemokine and TLR signalling pathways, were not significant in this analysis. By contrast, T cell receptor signalling and sulphur metabolism pathways were significantly enriched for immune-related genes but not for coeliac disease genes. Identifying further risk genes will most likely allow us to distinguish the higher-specificity pathways that are the genuine drivers for immune-related diseases from those that are specific to particular conditions.

Concluding remarks and future perspectives

At this time, 39 nonHLA genes are known to contribute to the susceptibility for coeliac disease. Although the findings account for a rather modest amount (~5%) of the total genetic risk, these genetic studies have expanded our understanding of the biology of coeliac disease and broadened the repertoire of immune pathways driving disease development. Genetic findings have now not only confirmed the well-established role of the adaptive immune response but have also indicated a clear role for the innate immune system, in particular by identifying genes involved in NF-κB signalling and viral infection. The genetic findings are now awaiting functional follow-up studies. One of the major remaining challenges is to pinpoint the true causal risk variants and to elucidate the consequences of these variants on RNA or proteins.

A GWAS comprising ~15,000 subjects was required to identify the loci that explain only 5% of the genetic variation. Hence, future studies will require much larger cohorts to uncover the remaining common variants with small effect sizes[57]. Analysis of 100,000 individuals might detect 10–15% of the genetic variance underlying coeliac disease[58]. It might be difficult to collate such a large collection of patient samples, but because many genes are shared with other diseases the joint analysis of multiple diseases might be one way forward.

Along with common variants, rare variants with a frequency of <10% are also likely to contribute to the missing heritability. Detecting rare variants will require different techniques, such as whole-exome or whole-genome sequencing[59]. Furthermore, structural variations, gene-gene interactions and epigenetic signatures are also likely to contribute to the genetic background of coeliac disease.

Because we have observed a large amount of shared genetics between coeliac disease and other immune-related traits, it is possible that the genetic background of immune diseases is generally shared to a greater degree than anticipated, and it is possible that the unique type of trigger determines whether coeliac disease develops rather than another disease. The roles of exogenous factors, such as infectious disease or gut microbiota, also need to be elucidated to better understand the interaction between the genetic background and environmental triggers. Once this information is available, a more complete picture of the disease pathogenesis will emerge. Only then can we begin to understand the heterogeneous clinical presentation of this disease and to translate our knowledge to benefit individual patients.

Acknowledgements
We thank Jihane Romanos, Cleo van Diemen, Alexandra Zhernakova and Jackie Senior for critically reading the manuscript. Our research is supported by grants from the Coeliac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009 to C.W.), the Netherlands Organisation for Scientific Research (NWO-VICI grant 918.66.620 to C.W.) and grants from the Wellcome Trust, Juvenile Diabetes Research Foundation International and Coeliac UK.

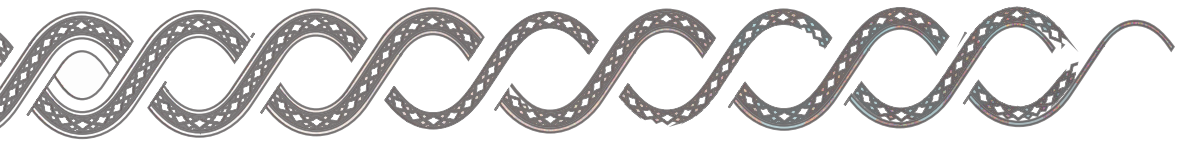
The authors have no conflicts of interest to report.

References

1 **Maki, M**. et al. (2003) Prevalence of celiac disease among children in Finland. N. Engl. J. Med. 348, 2517–2524
2 **D'Amico, M.A**. et al. (2005) Presentation of pediatric celiac disease in the United States: prominent effect of breastfeeding. Clin. Pediatr. 44, 249–258
3 **Rampertab, S.D.** et al. (2006) Trends in the presentation of celiac disease. Am. J. Med. 119, 355
4**Green,P.H**.(2005)The many faces of celiac disease: clinical presentation of celiac disease in the adult population. Gastroenterology 128, S74– 78
5 **Al-toma, A**. et al. (2007) Survival in refractory coeliac disease and enteropathy-associated T-cell lymphoma: retrospective evaluation of single-centre experience. Gut 56, 1373–1378
6 **Biagi,F. and Corazza, G.R.**(2010). Mortality in celiac disease.Nat.Rev. Gastroenterol. Hepatol. 7, 158–162
7 **Rubio-Tapia, A**. et al. (2009) Increased prevalence and mortality in undiagnosed celiac disease. Gastroenterology 137, 88–93
8 **Godfrey, J.D.** et al. (2010) Morbidity and mortality among older individuals with undiagnosed celiac disease. Gastroenterology 139, 763–769
9 **Molberg, O**. et al. (1998) Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. Nat. Med. 4, 713–717
10 **Jabri, B.** and Sollid, L.M. (2009) Tissue-mediated control of immunopathology in coeliac disease. Nat. Rev. Immunol. 9, 858–870
11 **Sollid, L.M**. et al. (1989) Evidence for a primary association of celiac disease to a particular HLA-DQ (alpha)/(beta) heterodimer. J. Exp. Med. 169, 345–350
12 **Spurkland, A.** et al. (1992) HLA-DR and -DQ genotypes of celiac disease patients serologically typed to be non-DR3 or non-DR5/7. Hum. Immunol. 35, 188–192
13 **Arentz-Hansen, H.** et al. (2000) The intestinal T cell response to (alpha)-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase. J. Exp. Med. 191, 603–612
14 **Nilsen, E.M.** et al. (1998) Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. Gastroenterology 115, 551–563
15 **Green, P.H. and Cellier, C.** (2007) Celiac disease. N. Engl. J. Med. 357, 1731–1743
16 **Meresse, B.** et al. (2004) Coordinated induction by IL15 of a TCR- independent NKG2D signaling pathway converts CTL into lymphokine-activated killer cells in celiac disease. Immunity 21, 357–366
17 **Dieterich, W.** et al. (1997) Identification of tissue transglutaminase as the autoantigen of celiac disease. Nat. Med. 3, 797–801
18 **Tye-Din, J.A.** et al. (2010) Comprehensive, quantitative mapping of T cell epitopes in gluten in celiac disease. Sci. Transl. Med. 2, 41ra51
19 **Dubois, P.C.A.** et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295–302
20 **Price, P.** et al. (1999) The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol. Rev. 167, 257–274
21 **Nistico, L.** et al. (2006) Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55, 803–804
22 **Van Heel, D.A.** et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827–829
23 **Hunt, K.A.** et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40, 395–402
24 **Trynka, G**. et al. (2009) Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-(kappa)B signalling. Gut 58, 1078–1083
25 **Smyth, D.J.** et al. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac diseasse. N. Engl. J.Med 359, 2767-2777
26 **Garner, C.P.** et al. (2009) Replication of celiac disease UK genome-wide association study results in a US population. Hum. Mol. Genet. 18, 4219–4225
27 **Nejentsev, S.** et al. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324, 387– 389

28 **Minamoto, S.** et al. (1997) Cloning and functional analysis of new members of STAT induced STAT inhibitor (SSI) family: SSI-2 and SSI-3. Biochem. Biophys. Res. Commun. 237, 79–83
29 **Zhernakova, A.** et al. (2010) Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am. J. Hum. Genet. 86, 970–977
30 **Bauquet, A.T.** et al. (2009) The costimulatory molecule ICOS regulates the expression of c-Maf and IL-21 in the development of follicular T helper cells and TH -17 cells. Nat. Immunol. 10, 167–175
31 **Yoshinaga, S.K.** et al. (2000) Characterization of a new human B7- related protein: B7RP-1 is the ligand to the co-stimulatory protein ICOS. Int. Immunol. 12, 1439–1447
32 **Minguet, S.** et al. (2008) The extracellular part of (zeta) is buried in the T cell antigen receptor complex. Immunol. Lett. 116, 203–210
33 **Bettelli, E**. et al. (2008) Induction and effector functions of TH17 cells. Nature 453, 1051–1057
34 **Fina, D.** et al. (2008) Regulation of gut inflammation and Th17 cell response by interleukin-21. Gastroenterology 134, 1038–1048
35 **Yamanouchi, J.** et al. (2007) Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity. Nat. Genet. 39, 329–337
36 **Klein, L.** et al. (2009) Antigen presentation in the thymus for positive selection and central tolerance induction. Nat. Rev. Immunol. 9, 833–844
37 **Allen, P.M.** (2009) Themis imposes new law and order on positive selection. Nat. Immunol. 10, 805–806
38 **Zamisch, M.** et al. (2009) The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8 T cell differentiation in the thymus. J. Exp. Med. 206, 2685–2699
39 **Wang, J. and Fu, Y.X.** (2003) LIGHT (a cellular ligand for herpes virus entry mediator and lymphotoxin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC self-peptide. J. Immunol. 170, 3986–3993
40 **Londei, M.** et al. (2005) Gliadin as a stimulator of innate responses in
celiac disease. Mol. Immunol. 42, 913–918
41 **Honda, K. and Taniguchi, T.** (2006) IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. Nat. Rev. Immunol. 6, 644–658
42 **Miller, S.D.** et al. (1997) Persistent infection with Theiler's virus leads to CNS autoimmunity via epitope spreading. Nat. Med. 3, 1133–1136
43 **Munz, C.** et al. (2009) Antiviral immune responses: triggers of or triggered by autoimmunity? Nat. Rev. Immunol. 9, 246–258
44 **Stene, L.C.** et al. (2006) Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am. J. Gastroenterol. 101, 2333–2340
45 **Perkins, N.D.** (2007) Integrating cell-signalling pathways with NF- kappaB and IKK function. Nat. Rev. Mol. Cell Biol. 8, 49–62
46 **Romanos, J. and Wijmenga, C.** (2010) Predicting susceptibility to celiac disease by genetic risk profiling. Ann. Gastroenterol. Hepatol. 1, 11–18
47 **Romanos, J. et al.** (2009) Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology 137, 834–840
48 **Norris, J.M.** et al. (2005) Risk of celiac disease autoimmunity and timing of gluten introduction in the diet of infants at increased risk of disease. J. Am. Med. Assoc. 293, 2343–2351
49 **Zhernakova, A.** et al. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet. 10, 43–55
50 **Coenen, M.J.H.** et al. (2009) Common and different genetic background for rheumatoid arthritis and coeliac disease. Hum. Mol. Genet. 18, 4195–4203
51 **Somers, E.C.** et al. (2006) Autoimmune diseases co-occurring within individuals and within families: a systematic review. Epidemiology 17, 202–217
52 **Eaton, W.W.** et al. (2007) Epidemiology of autoimmune diseases in Denmark. J. Autoimmun. 29, 1–9
53 **Barera, G.** et al. (2002) Occurrence of celiac disease after onset of type 1 diabetes: a 6-year prospective longitudinal study. Pediatrics 109, 833–838
54 **Hindorff, L.A.** et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. 106, 9362–9367
55 **Elbers, C.C.** et al. (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet. Epidemiol. 33, 419–431
56 **Fasano, A.** (2005) Clinical presentation of celiac disease in the pediatric population. Gastroenterology 128, S68–73
57 **Park, J.H.** et al. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. 42, 570–575
58 **Frazer, K.A.** et al. (2009) Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. 10, 241–251
59 **Cirulli, E.T. and Goldstein, D.B.** (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. 11, 415–425
60 **Manolio, T.A.** (2010) Genomewide association studies and assessment of the risk of disease. N. Engl. J. Med. 363, 166–176
61 **Huang and da, W.** et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44–57
62 **Kottgen, A.** et al. (2010) New loci associated with kidney function and chronic kidney disease. Nat. Genet. 42, 376–384
63 **Levy, D.** et al. (2009) Genome-wide association study of blood pressure and hypertension. Nat. Genet. 41, 677–687
64 **Newton-Cheh, C.** et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. Nat. Genet. 41, 666–676
65 **Cui, R.** et al. (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. Gastroenterology 137, 1768–1775
66 **Ganesh, S.K.** et al. (2009) Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat. Genet. 41, 1191–1198

67 **Kamatani, Y.** et al. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet. 42, 210–215
68 **Gudbjartsson, D.F.** et al. (2009) Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat. Genet.
41, 342–347
69 **Barrett, J.C.** et al. (2009) Genome-wide association study and meta- analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41, 703–707
70 **Imielinski, M.** et al. (2009) Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat. Genet. 41, 1335–1340
71 **De Jager, P.L.** et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple
sclerosis susceptibility loci. Nat. Genet. 41, 776–782
72 **Parkes, M.** et al. (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat. Genet. 39, 830–832
73 **Barrett, J.C.** et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. 40, 955– 962
74 **Wellcome Trust Case Control Consortium** (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678
75 **Todd, J.A.** et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet. 39, 857–864
76 **Cooper, J.D.** et al. (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat. Genet. 40, 1399–1401
77 **Gregersen, P.K.** et al. (2009) REL, encoding a member of the NF- kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. Nat. Genet. 41, 820–823
78 **Radstake, T.R.** et al. (2010) Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nat. Genet. 42, 426–429
79 **Yang, W.** et al. (2010) Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. PLoS Genet. 6, e1000841
80 **Han, J.W.** et al. (2009) Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. 41, 1234–1237
81 **Raychaudhuri, S.** et al. (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet. 41, 1313–1318
82 **McGovern, D.P.** et al. (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat. Genet. 42, 332-337
83 **Reveille, J.D.** et al (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat. Genet. 42,123-127
84 **Plenge, R.M.** et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat. Genet. 39, 1477–1482
85 **Stein, J.L.** et al. (2010) Voxelwise genome-wide association study (vGWAS). Neuroimage 15, 1160–1174
86 **Han, J.** et al. (2008) A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet. 4, e1000074
87 **Di Bernardo, M.C.** et al. (2008) A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat. Genet. 40, 1204–1210
88 **Sulem, P.** et al. (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat. Genet. 39, 1443–1452
89 **Nan, H.** et al. (2009) Genome-wide association study of tanning phenotype in a population of European ancestry. J. Invest. Dermatol. 129, 2250–2257
90 **Grant, S.F.** et al. (2009) Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes. Diabetes 58, 290–295
91 **Simon-Sanchez, J.** et al. (2009) Genome-wide association study reveals genetic risk underlying parkinson's diseasse. Nat. Genet. 41, 1308-1312
92 **Gieger, C.** et al. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet. 4, e1000282
93 **Jin, Y.** et al. (2010) Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. N. Engl. J. Med. 362, 1686–1697
94 **Codd, V.** et al. (2010) Common variants near TERC are associated with mean telomere length. Nat. Genet. 42, 197–199



Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease

Nat Genet in press

Gosia Trynka*, Karen A Hunt*, Nicholas A Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F Bakker, Maria Teresa Bardella, Leena Bhaw-Rosun, Gemma Castillejo, Emilio G de la Concha, Rodrigo Coutinho de Almeida, Kerith-Rae M Dias, Cleo C van Diemen, Patrick CA Dubois, Richard H Duerr, Sarah Edkins, Lude Franke, Karin Fransen, Javier Gutierrez, Graham AR Heap, Barbara Hrdlickova, Sarah Hunt, Leticia Plaza Izurieta, Valentina Izzo, Leo AB Joosten, Cordelia Langford, Maria Cristina Mazzilli, Charles A Mein, Vandana Midah, Mitja Mitrovic, Barbara Mora, Marinita Morelli, Sarah Nutland, Concepción Núñez, Suna Onengut-Gumuscu, Kerra Pearce, Mathieu Platteel, Isabel Polanco, Simon Potter, Carmen Ribes-Koninckx, Isis Ricaño-Ponce, Stephen S Rich, Anna Rybak, José Luis Santiago, Sabyasachi Senapati, Ajit Sood, Hania Szajewska, Riccardo Troncone, Jezabel Varadé, Chris Wallace, Victorien M Wolters, Alexandra Zhernakova, CEGEC (Spanish Consortium on the Genetics of Coeliac Disease), PreventCD Study Group, Wellcome Trust Case Control Consortium, BK Thelma, Bozena Cukrowska, Elena Urcelay, Jose Ramon Bilbao, M Luisa Mearin, Donatella Barisani, Jeffrey C Barrett, Vincent Plagnol, Panos Deloukas, Cisca Wijmenga, David A van Heel

*equal contribution

ABSTRACT
We densely genotyped, using 1000 Genomes Project pilot CEU and additional re-sequencing study variants, 183 reported immune-mediated disease non-*HLA* risk loci in 12,041 celiac disease cases and 12,228 controls. We identified 13 new celiac disease risk loci at genome wide significance, bringing the total number of known loci (including *HLA*) to 40. Multiple independent association signals are found at over a third of these loci, attributable to a combination of common, low frequency, and rare genetic variants. In comparison with previously available data such as HapMap3, our dense genotyping in a large sample size provided increased resolution of the pattern of linkage disequilibrium, and suggested localization of many signals to finer scale regions. In particular, 29 of 54 fine-mapped signals appeared localized to specific single genes - and in some instances to gene regulatory elements. We define a complex genetic architecture of risk regions, and refine risk signals, providing a next step towards elucidating causal disease mechanisms.

INTRODUCTION

Celiac disease is a common complex chronic immune-mediated disease with seroprevalence of ~1%[1,2] in individuals of white European origin. A T-cell mediated small intestinal immune response is generated against gliadin fragments from wheat, rye and barley cereal proteins leading to villous atrophy. Its aetiology is poorly understood. Association with *HLA* variants was first shown in 1972, and predisposing *HLA-DQ2* and *-DQ8* sub-types are necessary but not sufficient to cause disease. Recent genome wide association studies (GWAS) have identified a further 26 non-*HLA* risk loci[3-6]. Many of these loci are also associated with other autoimmune or chronic immune-mediated diseases (albeit sometimes different markers and effect directions[7]), with particular overlap observed between celiac disease, type 1 diabetes[8] and rheumatoid arthritis[9].

Currently unanswered questions regarding the genetic predisposition to celiac disease, which are also relevant for other immune-mediated diseases, include explaining the remaining major fraction of heritability, including rare and additional common risk variants; and identification of causal variants and causal genes (or at least more finely localizing the risk signal). The Immunochip Consortium[10] developed to explore these questions, taking advantage of emerging comprehensive common, low frequency, and rare variation datasets, and of a commercial offer of much lower per-sample custom genotyping costs for a very large project comprised of related diseases.

The Immunochip, a custom Illumina Infinium HD array, was designed to densely genotype, using 1000Genomes and any other available disease specific resequencing data, immune-mediated disease loci identified by common variant GWAS. The 1000 Genomes Project pilot CEU low-coverage whole genome sequencing dataset captures 95% of variants of MAF=0.05, and although underpowered to comprehensively detect variants of rarer allele frequency, still identifies 60% of variants of MAF=0.02, and 30% of variants of MAF=0.01[11]. The Consortium selected 186 distinct loci containing markers meeting genome wide significance criteria ($P<5x10^{-8}$) from twelve such diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis). All 1000 Genomes Project low-coverage pilot CEU population sample variants[11] (Sept 2009 release) within 0.1cM (HapMap3 CEU) recombination blocks around each GWAS region lead marker were submitted for array design. No filtering on correlated variants (linkage disequilibrium) was applied. Further case and control regional resequencing data were submitted by several groups (Online Methods, Supplementary Note), as well as a small proportion of investigator-specific undisclosed content including intermediate-significance GWAS results.

Most GWAS have been performed using common SNPs (typical minor allele frequency (MAF) >5%), further selected for low inter-marker correlation and/or even genomic spacing. In contrast to GWAS, the Immunochip presents a comprehensive in-depth opportunity to dissect the architecture of both rare and common genetic variation, at immuno-biologically relevant genomic regions, in human diseases. Due to the presence in our final Immunochip dataset of the majority of 1000 Genomes Project pilot CEU polymorphic genetic variants (and additional resequencing at some loci), the true causal variants from many risk loci may have been directly genotyped and analysed.

RESULTS

A total of 207,728 variants were submitted for Immunochip assay design and 196,524 passed manufacturing quality control at Illumina. After extensive and stringent data quality control (Online Methods), we analysed a near-complete dataset (overall 0.008% missing genotype calls)

comprising 12,041 celiac disease cases and 12,228 controls (from 7 geographic regions, Table 1) and 139,553 polymorphic (defined here as ≥2 observed genotype groups) markers. 634 biallelic SNPs were assayed in duplicate, at these we observed 189 of 15,384,884 (0.0012%) genotype calls to be discordant. Considering the intended 207,728 variants submitted for design, and an observed ~9.1% non-polymorphic rate in our post-quality control data, we estimate we have high quality genotype data on ~74% of the complete 1000 Genomes Project pilot CEU true polymorphic variant set at the fine-mapped regions.

We observed that 36 of the 183 non-HLA immune-mediated disease loci selected for Immunochip dense 1000Genomes-based genotyping achieved genome-wide significance ($P<5x10^{-8}$) for celiac disease in either the current study or our previous GWAS[5] (summary association statistics for all markers are available in T1DBase). All variants reaching genome wide significance were common (MAF >5%). We also observed marked enrichment for intermediate significance level celiac disease association signals (e.g. rs6691768, *NFIA* locus, P= $5.3x10^{-8}$) at a proportion of the remaining 147 dense-genotyped non-celiac autoimmune disease regions (Supplementary Figure 1). Variants from 3 dense-genotyped regions selected on Immunochip for a non-immune-mediated trait (bipolar disorder) showed no excess of association signals (Supplementary Figure 1).

We identified 13 new celiac risk loci ($P<5x10^{-8}$, Figure 1, Table 2, Supplementary Figure 2), 10 of which were from immune-mediated disease loci selected for Immunochip dense 1000Genomes-based genotyping. Several of these new loci were reported at lesser significance levels in our previous studies[5,9], and almost all have been reported in at least one other immune-mediated disease. These, with *HLA*, bring to 40 the total number of reported (current and/or previous study[5], which had an overlapping but slightly different sample set) genome wide significant celiac disease loci. Most contain candidate genes of immunological function, consistent with our previous findings at celiac disease loci[3-5].

Effect sizes (odds ratios, inverting protective effects) for the most significant marker per locus were median 1.155 (range 1.124 – 1.360) for the top signals from 26 non-HLA loci measured using Illumina Hap300/Hap550-chip linkage disequilibrium-pruned tag SNPs in our 2010 celiac disease GWAS[5] and median 1.166 (range 1.087 – 1.408) for the corresponding most significant marker (for the same signal) per locus in the current high density fine-mapping Immunochip dataset (Wilcoxon test P=0.75, Supplementary Table 1). Although we observe no difference in effect sizes between GWAS lead SNPs and subsequent fine-mapped signals, we note that case resequencing in the current Immunochip dataset is limited (see also Discussion).

In all, we report 57 independent coeliac disease association signals (Table 2) from 39 separate loci, of which 18 (32%) were not efficiently ($r^2>0.9$, Supplementary Table 2) tagged by our previous GWAS[5] (Illumina Hap550, post quality control dataset) markers.

Multiple independent common and rare variant signals

In contrast to most GWAS chips, the Immunochip contains a substantial proportion of lower MAF polymorphic variants. Of 139,553 variants in our 11,837 European-origin controls, 24,661 variants are low frequency (defined[11] as MAF 5% to 0.5%) and a further 22,941 variants are rare (MAF<0.5%). We investigated the possibility of multiple independently associated variants (of all allele frequencies) at each locus, using stepwise logistic regression conditioning on the most significant variant at the locus (Online Methods, Supplementary Table 3). This analysis can be sensitive to genotype miscalling and missing data[12], hence our use of extremely rigorous quality control measures for the dataset and manual inspection of genotype clusters for all reported markers.

Table 1 Sample collections

Population sample	Celiac cases	Controls
UK	7728	8274[b]
The Netherlands	1123	1147
Poland	505	533
Spain - CEGEC [a]	545	308
Spain - Madrid[a]	537	320
Italy - Rome, Milan, Naples	1374	1255
India - Punjab	229	391
Total	12041	12228

[a]The two Spanish population samples were considered separately due to genotyping in different laboratories.
[b]5430 UK 1958 Birth Cohort participants, and 2844 UK Blood Services-Common Controls. Each of the collections from the UK, Netherlands, Poland, Spain (Madrid) and Italy contained essentially the same sample set as our 2010 celiac disease GWAS[5], with now substantial additional samples from the UK and Netherlands and exclusion of amplified DNA samples from the Spanish collections. The Indian collection has not previously been studied. Our 2010 GWAS contained several collections not studied here.

We observed two or more independent signals at 13 of 36 high-density genotyped non-HLA loci (Figure 2). Four of these loci each had three independent signals (*STAT4*, the chromosome 3 *CCR* region, *IL12A*, *SOCS1/PRM1/PRM2*, Table 2). Low frequency and/or rare variant signals were seen at four separate loci (*RGS1*, *CD28/CTLA4/ICOS*, *SOCS1/PRM1/PRM2*, *PTPN2*). Notably, the strongest effect (OR 1.70) was seen at the rare variant imm_16_11281298 (*SOCS1/PRM1/PRM2* locus) with genotype counts (AA/AG/GG) of 1/136/11904 (MAF 0.57%) in all celiac cases and 0/91/12136 (MAF 0.37%) in all controls (detailed genotype count and allele frequency data for top signals by collection are shown in Supplementary Table 4).

We next performed haplotype analysis on all loci with multiple independent signals, to investigate whether the multiple signals were due to multiple causal effects or a single effect best tagged by several variants. For all but one locus (*PTPN2*) the haplotype association tests (not shown) were of similar significance to the single SNP association tests, suggesting that for each signal we have genotyped either the causal variant, or markers very strongly correlated with it. These findings contrast with those from a recent resequencing study[13], probably because of the much greater variant density of our study. However, at the *PTPN2* locus, the imm_18_12833137(T) + ccc-18-12847758-G-A(G) haplotype was considerably more associated (P=4.8x10^{-14}, OR 0.84) than either SNP alone (imm_18_12833137 P=1.9x10^{-10}; ccc-18-12847758-G-A P= 0.0008).

Interestingly at the *SOCS1* locus, the third independent signal imm_16_11292457 shows association only after conditioning on the two other signals (P=2.0x10^{-4}) but not in the single SNP non-conditioned association analysis (P=0.15). Further inspection revealed the protective imm_16_11292457(A) allele to be correlated (in linkage disequilibrium) with the risk (A) allele of the first signal imm_16_11268703, thus although there are indeed three independent signals, the effect of the third signal is only revealed after conditioning on the first. A similar statistical effect (Simpson's paradox) was recently shown at a Parkinson's disease locus[14].

Fine-mapping to localize causal signals

GWAS signals are typically reported within relatively large linkage disequilibrium blocks. We tested whether our much denser genotyping strategy would allow finer-scale localization, and the pinpointing of association signals. We found that markers strongly correlated (r^2>0.9) with the most significant independent variant clustered together, and defined regions that are a median 12.5x smaller than the relevant HapMap3 CEU 0.1cM linkage disequilibrium blocks (Table 2, Figure 2, Supplementary Figure 2). Localization was highly successful for some regions (e.g. *PTPRK, TAGAP*), but not possible at others (e.g. *IL2-IL21*). At many loci, the

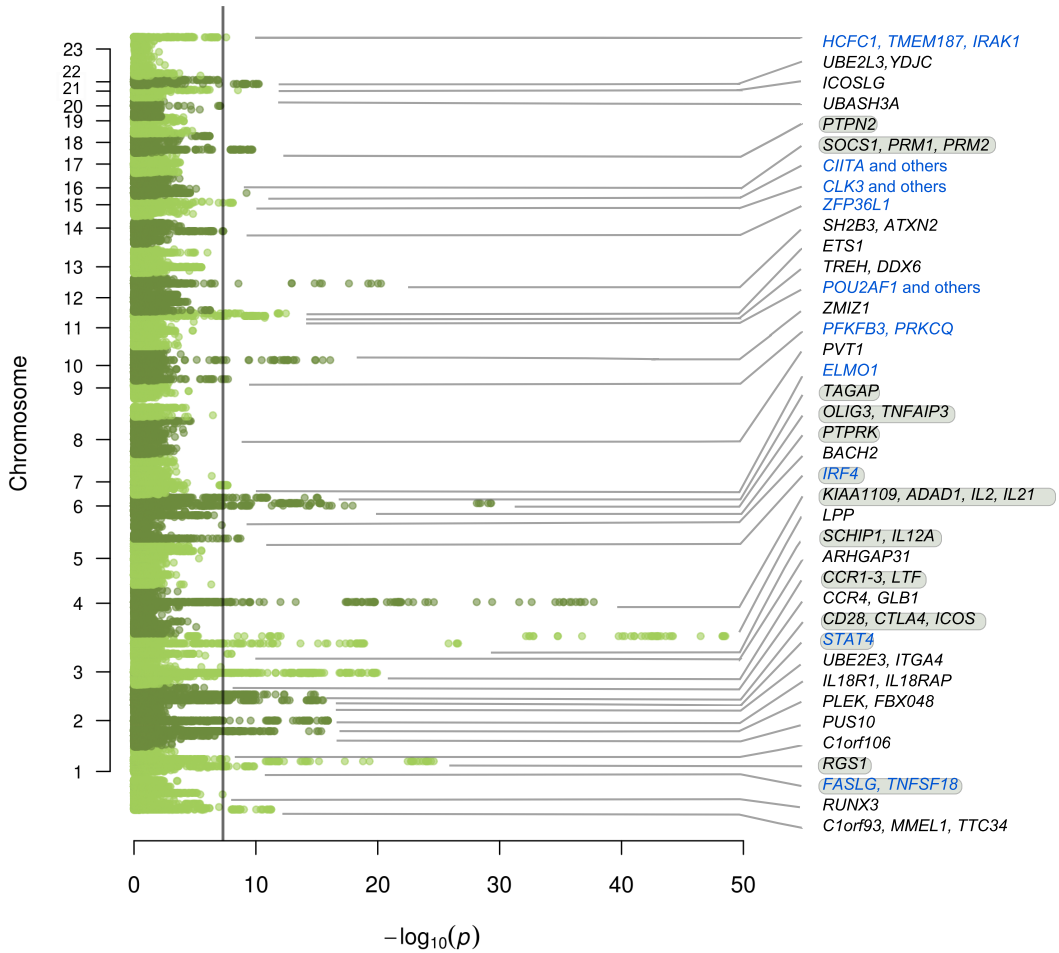


Figure 1 Manhattan plot of association statistics for known and novel celiac disease risk loci Novel loci indicated in blue, loci with multiple signals indicated with grey highlight. Significance threshold drawn at $P=5x10^{-8}$.

localized regions comprised only a handful of markers in close physical proximity.

Considering the 36 high density genotyped loci, we have localized to a single gene 29 of the total 54 independent non-*HLA* signals (Table 2, Supplementary Figure 2). We identified all markers strongly correlated ($r^2>0.9$) with the independent non-*HLA* variants reported in our analyses (from Table 2), and on functional annotation (Supplementary Table 2) identified only a handful of markers in exonic regions and of these only three are protein altering variants (nsSNPs: imm_1_2516606 (*MMEL1*), imm_12_110368991(*SH2B3*),1kg_X_152937386 (*IRAK1*). In contrast, a number of signals appeared to be more finely localized around the transcription start site of specific genes (which we defined as the first exon, and 10kb 5′ of the first exon), including signals at *RUNX3*, *RGS1*, *ETS1*, *TAGAP*, *ZFP36L1*; and around the 3′ UTR region (and 10kb 3′) including signals at *IRF4*, *PTPRK* and *ICOSLG*.

Overlap between multiple independent signal regions was seen at some loci (Figure 2), suggesting that causal variants might be functioning through a shared mechanism e.g. within a 2kb region of the *PTPRK* 3′ UTR; within a 11kb region 5′ of *IL12A*; or within a 28kb region

of *TNFAIP3*. In contrast, multiple independent signals were observed that spread between the three immune genes of the *CD28*/*CTLA4*/*ICOS* region.

DISCUSSION

We show that fine mapping of GWAS regions using dense resequencing data, e.g. (as here) from the 1000Genomes project, is feasible and generates substantial additional information at many loci. We identify a complex architecture of multiple common and rare genetic risk variants at around a third of the now 40 proven celiac disease loci. The design of our study has allowed us to find many more such complex regions than the ~10% with multiple signals seen in our previous study[5] and a recent large GWAS for human height[15]. It seems probable that if larger sample sizes than in the current study were to be tested, additional loci might be shown to have a similarly rich multiple risk variant architecture. Multiple independent risk signals for celiac disease have also long been known in the *HLA* region[16]. Our success in celiac disease might be partly due to the extensive selective pressures for haplotypic diversity that have taken place at immune gene loci[17]. Previous studies have reported independently associated common and rare variants at individual loci for a handful of phenotypes e.g. fetal haemoglobin[13], sick sinus syndrome[18], Crohn's disease[19], hypertriglyceridemia[20]. To the best of our knowledge, ours is the first study to have comprehensively surveyed the genetic architecture of all known risk loci for a trait.

In part, our identification of rare variants at risk regions relies on the prior discovery of a genome-wide significant common variant association signal at each locus. This then permits a per-locus rather than genome-wide multiple testing correction when searching for additional independent association signals. Only particularly strong rare variant signals would, on their own, generate significance levels reaching the genome-wide threshold typically used in GWAS studies ($P<5x10^{-8}$). Alternative methods, such as collapsing rare variant signals across a gene or functional categories of genes have therefore been suggested as approaches to the same problem[21]. Although a rare variant may have occurred on a recent haplotypic background, and thus show linkage disequilibrium at substantially longer range than common variants, we deliberately restricted our search to around the common variant linkage disequilibrium blocks as to do otherwise would have incurred a considerably greater penalty from multiple testing. Therefore, although our study provides considerable encouragement for exome and whole genome sequencing efforts aimed at identifying rare risk variants (not necessarily restricted to GWAS loci) in common complex diseases, it further highlights the statistical challenges of establishing rare variant associations.

We used a dense genotyping strategy and stepwise conditional association analysis, but did not identify any rare highly penetrant variants that might explain the genome-wide significant common SNP signals at any of the 39 loci. Our study does have limitations in this regard, particularly i) analysis restricted to 0.1cM linkage disequilibrium blocks; ii) the limited control resequencing sample size of the 1000 Genomes Project pilot CEU dataset; iii) the limited case resequencing sample size; and iv) case resequencing limited to three loci for celiac disease, and selected loci for other immune diseases. We observed a weak trend towards lower MAF (P=0.042, Wilcoxon test, Supplementary Table 1) for the best fine-mapping SNP (Immunochip experiment) versus the lead SNP from our 2010 tag SNP GWAS (measuring MAF in a subset of samples genotyped in both datasets). One signal showed substantially higher MAF (>25% change) on fine-mapping, four signals showed substantially lower MAF on fine mapping (Supplementary Table 1), yet all fine-mapping variants corresponding to lead GWAS SNPs remained common (MAF>0.10). We suggest that these changes in MAF upon fine-mapping of lead GWAS SNPs simply reflect more precise measurement of common frequency risk haplotypes. Although we cannot exclude

Table 2 Risk variant signals at genome-wide significant celiac disease loci.

Top variant (dbSNP130 id)	Chr	HapMap3 CEU LD block[b] positions (hg18) (n markers, size)	MAF[c]	P[d]	OR	Highly correlated (r^2>0.9) variants positions (hg18) (n markers, size)	Localization: protein coding genes (RefSeq track UCSC/ hg18)
rs4445406	1	2396747 - 2775531 (358, 379kb)	0.344	$5.4x10^{-12}$	0.87	2510162 - 2710035 (27, 200kb)	C1orf93, MMEL1, TTC34
rs72657048	1	25111876 - 25180863 (125, 69kb)	0.498	$3.8x10^{-6}$	0.92	25162321 - 25177139 (18, 15kb)	0 - 10kb 5′ & 1 st exon RUNX3
rs12068671	1	170917308 - 171207073 (355, 290kb)	0.185	$1.4x10^{-10}$	0.86	170940206 - 170948695 (11, 8kb)	35 - 43kb 5′ FASLG
signal 2 rs12142280	1	"	0.180	$8.3x10^{-9}$ d	0.87	171129607 – 171131275 (2, 2kb)	intergenic between FASLG and TNFSF18
rs1359062	1	190728935 - 190814664 (181, 86kb)	0.180	$2.5x10^{-25}$	0.77	190786488 - 190811722 (17, 25kb)	0 - 24kb 5′ & 1 st exon RGS1
signal 2 rs72734930	1	"	0.022	$3.7x10^{-4}$ d	1.23	190779182 (1)	32kb 5′ RGS1
rs10800746	1	199119734 - 199308949 (331, 189kb)	0.305	$2.6x10^{-8}$	0.89	199148015 (1)	9th intron C1orf106
rs13003464	2	60768233 - 61745913 (1047, 978kb)	0.388	$4.3x10^{-16}$	1.17	61040333 - 61058360 (3, 18kb)	exons 5-11 PUS10
rs10167650	2	68389757 - 68535760 (357, 146kb)	0.266	$1.3x10^{-4}$	0.92	68493221 - 68499064 (4, 6kb)	intergenic between PLEK and FBX048
rs990171	2	102221730 - 102573468 (894, 352kb)	0.225	$1.2x10^{-16}$	1.20	102338297 - 102459513 (45, 121kb)	IL18R1, IL18RAP
rs1018326	2	181502502 - 181972196 (898, 470kb)	0.418	$3.1x10^{-16}$	1.16	181708291 - 181803246 (24, 95kb)	intergenic between UBE2E 3 and ITGA4
rs6715106	2	191581798 - 191715979 (203, 134kb)	0.058	$8.4x10^{-9}$	0.79	191621279 - 191643278 (4, 22kb)	exons 6-14 STAT4
signal 2 rs6752770	2	"	0.296	$1.3x10^{-6}$ d	1.10	191681808 (1)	intron 3 STAT4
signal 3 rs12998748	2	"	0.119	$2.6x10^{-4}$ d	0.90	191656882 (1)	intron 3 STAT4
rs1980422	2	204154625 - 204524627 (642, 370kb)	0.233	$1.4x10^{-15}$	1.19	204318641 - 204320303 (2, 2kb)	intergenic between CD28 and CTLA4
signal 2 rs3407980	2	"	0.217	$1.6x10^{-5}$ d	0.91	204470572 – 204478299 (2, 8kb)	intergenic between CTLA4 and ICOS
signal 3 rs10207814	2	"	0.039	$1.3x10^{-4}$ d	1.20	204158521 - 204168206 (5, 10kb)	111 – 121 kb 5′ CD28
rs4678523	3	32895606 - 33063377 (260, 168 kb)	0.313	$2.4x10^{-7}$	1.11	33012725 - 33012756 (2, 31bp)	intergenic between CCR4 and GLB1
rs2097282	3	45904804 - 46625997 (1343, 721kb)	0.314	$1.1x10^{-20}$	1.20	46321275 - 46377631 (27, 56kb)	intergenic between CCR3 and CCR2
signal 2 rs7616215	3	"	0.361	$8.6x10^{-9}$ d	1.12	46162711 – 46180690 (2, 18kb)	38 – 55 kb 3′ CCR1
signal 3 rs60215663	3	"	0.070	$4.8x10^{-5}$ d	1.16	46458634 – 46480319 (7, 22kb)	exons 2-13 LTF (NM_002343.3)
rs61579022	3	120587671 - 120783345 (372, 196kb)	0.390	$9.9x10^{-9}$	1.11	120601187 - 120605968 (4, 5kb)	intron 10 ARHGAP31
[imm_3_161120372]	3	161065075 - 161237201 (423, 168kb)	0.111	$2.6x10^{-27}$	1.36	161112778 - 161147744 (4, 35kb)	intergenic between SCHIP1 and IL12A
signal 2 rs1353248	3	"	0.288	$9.8x10^{-9}$ d	0.88	161106253 (1)	intergenic between SCHIP1 and IL12A
signal 3 rs2561288	3	"	0.455	$8.1x10^{-8}$ d	1.12	161136316 – 161168494 (6, 32kb)	intergenic between SCHIP1 and IL12A
rs2030519	3	189552054 - 189622323 (142, 70kb)	0.486	$3.0x10^{-49}$	0.76	189587750 - 189602595 (8, 15kb)	intron 2 LPP
rs13132308	4	123192512 - 123784752 (1294, 592kb)	0.166	$1.9x10^{-38}$	0.71	123269042 - 123770564 (11, 502kb)	multiple genes (KIAA1109, ADAD1, IL2, IL21)
signal 2 rs62323881	4	"	0.073	$8.6x10^{-5}$ d	1.15	123257527 – 123722990 (87, 465kb)	multiple genes (KIAA1109, ADAD1, IL2, IL21)
rs1050976	6	315547 - 402748 (199, 87kb)	0.488	$1.8x10^{-9}$	0.89	353079 - 355417 (3, 2kb)	3′ UTR IRF4 (NM_002460.3)
signal 2 rs12203592	6	"	0.183	$2.6x10^{-4}$ d	0.91	341321 (1)	intron 4 IRF4 (NM_002460.3)
rs7753008	6	90863556 - 91096529 (341, 233kb)	0.380	$2.7x10^{-7}$	1.10	90866360 - 90875874 (5, 10kb)	intron 2 BACH2 (NM_001170794.1)
rs55743914	6	127993875 - 128382483 (572, 389kb)	0.239	$1.1x10^{-18}$	1.21	128332892 - 128335255 (2, 2kb)	PTPRK last exon, 3′UTR (NM_002844.3)
signal 2 rs72975916	6	"	0.150	$1.2x10^{-5}$ d	0.89	128307943 - 128339304 (15, 31kb)	PTPRK exons 28-30, 3′UTR , to 24kb 3′
rs17264332	6	137924568 - 138316778 (864, 392kb)	0.211	$5.0x10^{-30}$	1.29	138000928 - 138048197 (6, 47kb)	intergenic between OLIG3 and TNFAIP3
signal 2 [imm_6_138043754]	6	"	0.190	$2.1x10^{-7}$ d	0.88	138015797 – 138043754 (4, 28kb)	intergenic between OLIG3 and TNFAIP3
rs182429	6	159242314 - 159461818 (514, 220kb)	0.427	$8.5x10^{-16}$	1.16	159385965 - 159390046 (4, 4kb)	4kb 5′ and 5′ UTR TAGAP (NM_152133.1)
signal 2 rs1107943	6	"	0.071	$2.8x10^{-6}$ d	1.18	159418255 (1)	32kb 5′ TAGAP (NM_152133.1)
[1kg_7_37384979]	7	37330503 - 37406978 (213, 76kb)	0.101	$2.1x10^{-8}$	1.18	37366994 - 37404402 (31, 37kb)	intron 1 ELMO1
rs10808568	8	129211716 - 129368419 (400,157kb)	0.256	$2.2x10^{-5}$	0.91	129333242 - 129345888 (4, 13kb)	151 - 163kb 3' of PVT1
rs2387397	10	6428077 - 6585110 (411, 157kb)	0.229	$1.9x10^{-8}$	0.88	6430198 (1)	intergenic between PFKFB3 and PRKCQ

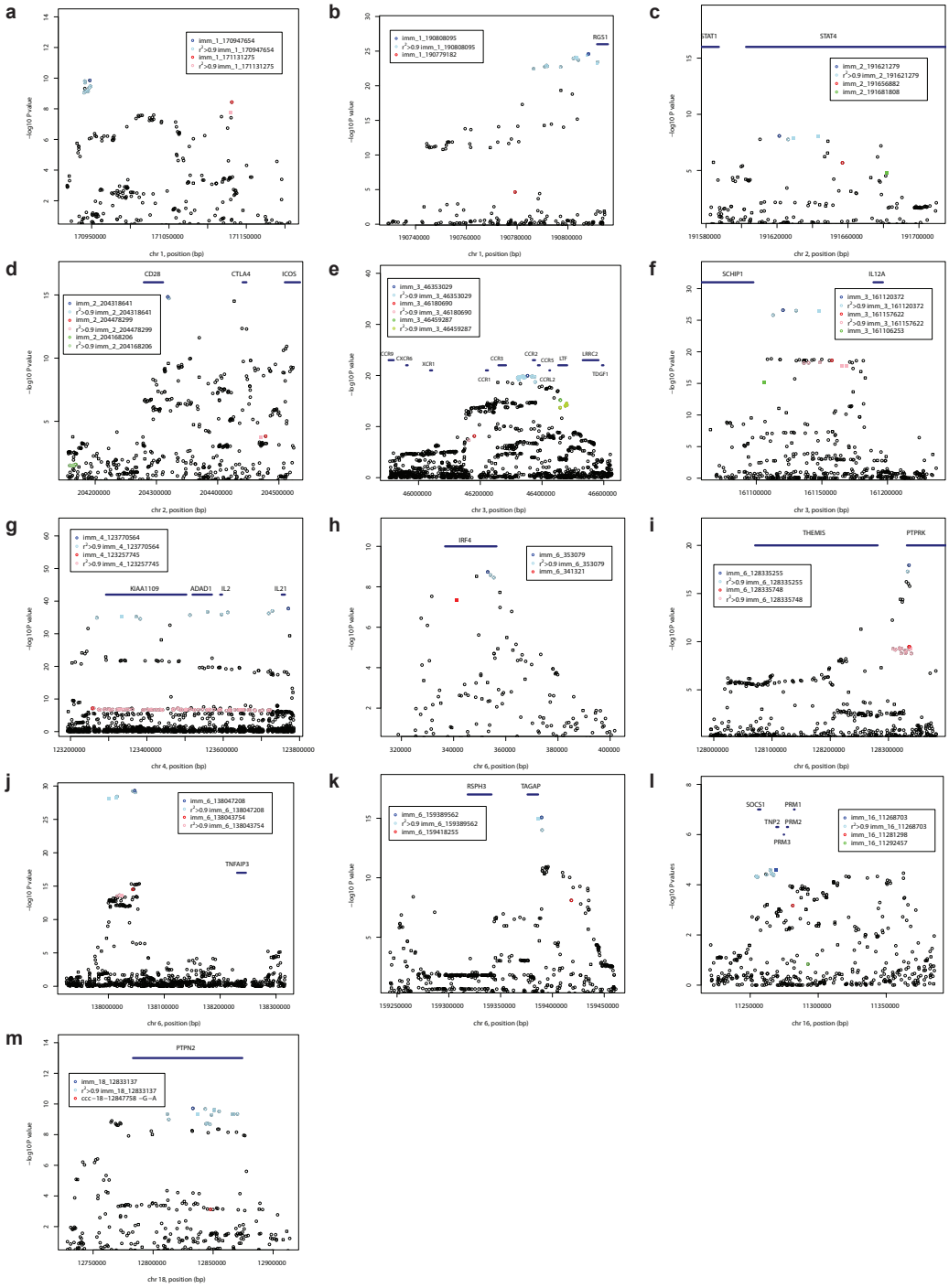
rs1250552	10	80690408 - 80774414 (223, 84kb)	0.470	$8.0x10^{-17}$	0.86	80728033 (1)	intron 14 ZMIZ1
rs7104791	11	110682429 - 110815769 (3, 133kb)	0.209	$1.9x10^{-11}$	1.16	not high-density genotyped	[region: POU2AF1 , C11orf93]
rs10892258	11	117847131 - 118270810 (466, 424kb)	0.237	$1.7x10^{-11}$	0.86	118080536 - 118085075 (5, 5kb)	intergenic between TREH and DDX6
rs61907765	11	127754640 - 127985723 (480, 231kb)	0.213	$3.4x10^{-13}$	1.18	127886184 - 127901948 (6, 16kb)	5kb 5' & 1 st exon ETS1 (NM_001162422.1)
rs3184504	12	110183529 - 111514870 (938, 1331kb)	0.488	$5.4x10^{-21}$	1.19	110368991 - 110492139 (4, 123kb)	5' UTR & exons 1 -3 $SH2B3$; exons 2-25 & 3' UTR ATXN2
rs11851414	14	68238574 - 68387815 (338, 149kb)	0.221	$4.7x10^{-8}$	1.13	68329159 - 68341722 (3, 13kb)	1kb 5' & 1st exon ZFP36L1
rs1378938	15	72397784 - 73270664 (23, 873kb)	0.278	$7.8x10^{-9}$	1.13	not high-density genotyped	[region inc. CLK3 , CSK and multiple genes]
rs6498114	16	10834038 - 10903351 (8, 69kb)	0.246	$5.8x10^{-10}$	1.14	not high-density genotyped	[region: CIITA]
rs243323	16	11220552 - 11385420 (446, 165kb)	0.300	$2.5x10^{-5}$	0.92	11254549 - 11268703 (12, 14kb)	11kb 5', all of SOCS1, 1kb 3'
signal 2 [imm_16_11281298]	16	"	0.004	$1.3x10^{-4}$ [d]	1.70	11281298 (1)	intergenic between PRM1 and PRM2
signal 3 rs9673543	16	"	0.169	$2.0x10^{-4}$ [d]	1.10	11292457 (1)	10kb 5' PRM1
rs11875687	18	12728413 - 12914117 (411, 186kb)	0.150	$1.9x10^{-10}$	1.17	12811903 - 12870206 (16, 58kb)	exons 2-5 PTPN2 (NM_080422.1)
signal 2 rs62097857	18	"	0.040	$5.2x10^{-5}$ [d]	1.20	12847758 (1)	intron 2 PTPN2 (NM_080422.1)
rs1893592	21	42683153 - 42760214 (226, 77kb)	0.282	$3.0x10^{-9}$	0.88	42728136 (1)	intron 9 UBASH3A (NM_018961)
rs58911644	21	44414408 - 44528088 (239, 114kb)	0.193	$6.2x10^{-7}$	0.89	44446245 - 44453549 (8, 7kb)	18 - 25kb 3' ICOSLG
rs4821124	22	20042414 - 20352005 (131, 310kb)	0.186	$5.7x10^{-11}$	1.16	20250903 - 20313260 (36, 62kb)	UBE2L3 , YDJC
rs13397	X	152825373 - 153043675 (88, 218kb)	0.133	$2.7x10^{-8}$	1.18	152872114 - 152937386 (4, 65kb)	HCFC1 , TMEM187 , IRAK1

Non-HLA loci meeting genome-wide significance ($P<5x10^{-8}$) in the current Immunochip dataset, or previous GWAS/replication dataset[5], are shown. Loci reported for the first time for celiac disease at genome wide significance are shown in bold in the Top variant column.

the possibility that a single high-penetrance lower-frequency variant explains most of the association signal at a locus, especially without more comprehensive case resequencing, we find no evidence in support this possibility in the current fine-mapping experiment. Nor can our stepwise selection procedure robustly refute the "synthetic association" hypothesis - in particular that a combination of multiple rare variants jointly explains the association signal[22] - although similarly we have not observed so far evidence supporting this possibility.

We established at genome wide significance 13 new loci for celiac disease, most of which have been reported previously at lesser significance or for another immune-mediated disease. The Illumina Hap550 chip (used in our 2010 GWAS) should have detected 10 of the 13 new loci, and in total 39 of the 57 independent non-HLA signals that we report. A current genotyping platform, the Illumina Omni2.5 chip would have detected 12 of the 13 new loci, and in total 50 of the 57 independent non-HLA signals that we report. Neither chip would have provided the finer scale localization of the Immunochip. The thirteen new loci contain many candidate genes of immunological function (P=0.0002 for enrichment of the Gene Ontology term "immune system process"[23]), in line with expectations from our previous studies. We also show evidence suggesting substantial additional signals at other immune-mediated disease loci, which lie beneath the genome wide significance reporting threshold applied to the current dataset. It is a point of debate whether such strict ($P<5x10^{-8}$) criteria should apply - a Bayesian analyst might apply a higher prior at a locus already reported in another immune-mediated disease. Alternatively, an Immunochip-wide P value with a Bonferroni correction for independent SNPs, as used recently for the Cardiochip custom genotyping project[24], of $P<1.9 \times 10^{-6}$ (Online Methods) would yield 16 additional celiac disease loci. These 16 loci also mostly contain immune system genes. An analysis of these currently intermediate significance signals would gain substantial additional power by a meta-analysis across the several hundred thousand samples from multiple immune-mediated disease collections presently being run on Immunochip.

Figure 2 Loci with multiple independent signals



Non-conditioned P values shown for loci with multiple independent signals (from **Table 2**). The most associated variant for a signal shown in bold colour, further variants in r^2>0.90 (calculated from the 24,249 sample Immunochip dataset) shown in normal colour. First signal coloured blue, second coloured red, third coloured green. Squares indicate markers present in our previous celiac disease GWAS post quality control dataset (Illumina Hap550)[5]

We found that our previous GWAS using tag SNPs gave very similar estimates of effect size to our current fine-mapping experiment (Supplementary Table 1), in contrast to a simulation study which suggested that GWAS markers often underestimate risk[14]. We have, however, found substantial evidence for multiple additional signals at known loci and report many new loci. In Europeans, the current 39 non-*HLA* loci now explain 13.7% of coeliac disease genetic variance (*HLA* accounts for a further ~40%). We also show a long tail of likely effects of weaker significance, which will explain substantial additional heritability.

Only one of the variants reported here was discovered by a disease-specific resequencing study: ccc-18-12847758-G-A (rs62097857), a marker identified by the WTCCC group's resequencing of Crohn's disease cases and controls (Supplementary Note) and also present in the Watson genome. We submitted for Immunochip ~4,000 variants from high throughput resequencing of pools of 80 celiac disease cases for extended genomic regions at three loci (*RGS1*, *IL12A*, *IL2-IL21*, Supplementary Note). These did not contribute additional signals over and above those obtained from the 1000 Genomes Project pilot CEU variants, although did contribute to increase the numbers of variants correlated with each signal (i.e the set of markers that likely contains the causal variant(s)) and more precisely define the bounds of the signal localization. We note that larger scale case resequencing (e.g. many hundreds of samples) would identify a rarer spectrum of variants than the current study, and has previously been used with success at selected genes and phenotypes.

The possibility of performing fine-scale mapping of GWAS regions using e.g. 1000 Genomes Project data has been discussed as a natural follow-on strategy for such studies[25,26] and has been recently used to identify risk variants in *APOL1* in African-Americans with renal disease[27]. Our current report is the first to test such a strategy on a large scale in a complex disease. At multiple regions, we were able to refine the signal to a handful of variants over a few kilobases or tens of kilobases, although some regions (e.g. *IL2-IL21*) were resistant to this approach presumably due to particularly strong linkage disequilibrium. Most GWAS publications report signals mapping to a "LD block" based on HapMap recombination rates (sample size, 60 CEU families). In our data, where we have both i) much denser genotyping than GWAS chips (mean 13.6x at celiac loci versus the Illumina Hap550 chip) and ii) nearly 25,000 genotyped samples for the linkage disequilibrium calculations, we are able to observe much finer scale recombination and more precisely estimate of the bounds of nominimal recombination intervals. Our findings are similar in terms of genotyping density and the resulting fine-mapped region size and lack of haplotype-specific effects to an earlier study of the *IL2RA* locus in type 1 diabetes[26]. At the majority of regions a tight block of highly correlated variants was seen, rather than a gradual decay of correlation (e.g. Figure 2 plots for *IL12A*, *PTPRK*). At many loci, we have now defined a handful of likely candidates to be the causal variant(s) to be taken forward into functional studies, although we may have missed candidate variants at some regions due to the sample size of the 1000 Genomes Project pilot CEU dataset (60 individuals), their status as controls, and our estimate that ~25% of these variants were excluded from our final dataset. These might be assessed by imputation methods[28], but our approach – particularly with regards to the more sensitive conditional regression analysis - has been to prefer the more accurate direct genotyping of all assayable variants. As and when much larger whole genome resequencing based reference datasets become available (e.g. the main 1000 Genomes Project), these might be used to impute into our Immunochip dataset, including substantially lower frequency variants[29]. We also investigated whether our use of multiple ethnic subgroups within Europe (e.g. southern European Spanish versus northern European UK) or the relatively small Indian collection contributed to fine mapping, and found that in most cases, the same degree of localization

was possible with just the UK collection alone (data not shown).

Our data suggest that most common risk variants might function by influencing regulatory regions, consistent with those previously reported in other immune-mediated diseases, and complex traits in general[11]. The exception is the *SH2B3* nsSNP imm_12_110368991 (rs3184504), reported in our 2008 celiac GWAS[4], which even with the fine-mapping of 938 polymorphic variants from the *SH2B3* region remains the strongest signal at this locus thus suggesting it may be the causal variant. The same variant has been associated with other immune diseases, and a functional immune phenotype[5]. Interestingly, we observed a common ~980bp intergenic deletion between *IL2* and *IL21* (DGV40686, accurately genotyped by Infinium assay with control MAF 7.3%) correlated with the second independent signal at this region, although we have no evidence to suggest causality.

Our fine-scale localization approach has identified likely causal genes at many loci, and at eight genes signals localized around the 5′ or 3′ regulatory regions. For example, at the *THEMIS*/*PTPRK* locus, two independently associated sets of variants cluster in the 3′ UTR of the *PTPRK* gene (one, imm_6_128332892/rs3190930 in a predicted binding site for miRNA hsa-miR-1910). *PTPRK*, a TGF-beta target gene, is involved in $CD4^+$ T cell development and a deletion mutation causes T helper deficiency in the LEC rat strain[30]. The signal at *TAGAP* lies within a 4kb region immediately 5′ of the transcription start site, presumably containing promoter elements. At *ETS1*, the signal comprises 6 variants overlapping the promoter and 1st exon of the T cell expressed isoform NM_001162422.1, and one of the variants (imm_11_127897147/rs61907765) has predicted regulatory potential and overlaps multiple transcription factor binding sites (UCSC GenomeBrowser ChipSeq and ESPERR tracks, Supplementary Table 2). Similarly interesting variants are observed in regulatory regions of *RUNX3* (imm_1_25165788/rs11249212), and *RGS1* (imm_1_190807644/rs1313292, imm_1_190811418/rs2984920) (Supplementary Table 2). Such an approach to identify the functional potential of risk variants was recently successful used to define a causal systemic lupus erythematosus *TNFAIP3* variant[31]. Although we have localized signals at many loci, and recent research suggests the likely causal gene is often located near the most strongly associated variant[15], only more detailed functional studies (e.g. transcription factor binding assays[31] and transcriptional activity assays of constructs with individual single nucleotide alterations at risk SNPs[32]), will prove precisely which gene variants might be causal.

We conclude that dense fine mapping of regions identified through GWAS studies can uncover a complex genetic architecture of independent common and rare variants, and often successfully localize risk variant signals to a small set of SNPs to be taken forward into functional assays. Denser fine mapping studies, utilising larger resequencing sample sizes from both cases and controls over broader regions, might provide further resolution of GWAS signals.

ONLINE METHODS

Subjects. Written informed consent was obtained from all subjects, with Ethics Committee / Institutional Review Board approval. All individuals, except the Indian population sample, are of European ancestry. DNA samples were from blood, lymphoblastoid cell lines or saliva.

Affected celiac individuals were diagnosed according to standard clinical criteria, compatible serology and in all cases small intestinal biopsy - most cases were diagnosed using the revised ESPGHAN criteria as a minimum requirement[34]. More specific requirements were: UK cases[3-5] (hospital outpatients, n=1145) required Marsh III stage intestinal biopsy (HLA-DQ2.5cis tag SNP rs2187668 MAF=0.4699); UK cases [4,5] (Coeliac

UK members, n=6583), self-reported diagnosis by intestinal biopsy (note the rs2187668 MAF=0.4803 was similar to UK hospital cases, versus combined UK controls MAF 0.1419); Italy (Milan) [5,35] and Polish[5] required Marsh III stage intestinal biopsy and positive endomysial/ tissue transglutaminase antibodies; Spain (CEGEC)[36] required at least Marsh II stage intestinal biopsy; Netherlands cases[5] required Marsh III stage intestinal biopsy, or Marsh II stage intestinal biopsy with compatible HLA-DQ type; India (Punjab) cases required Marsh III stage intestinal biopsy and strongly positive tissue transglutaminase antibodies; Italy (Naples, Rome) required abnormal intestinal biopsy and positive tissue transglutaminase antibodies[37].

The UK 1958 Birth Cohort and UK Blood Services-Common Controls are unselected population controls. Polish controls and Italian (Naples) controls excluded celiac serology positive samples. Spain (Madrid) controls were unselected blood donors and hospital employees. Spain (CEGEC), Italy (Rome), Indian (Punjab) controls were unselected blood donors. Italian controls (Milan) were unselected healthy individuals. Netherlands controls were unselected blood donors and population controls.

SNP selection: All 1000 Genomes Project low-coverage whole genome sequencing pilot CEU variants within 0.1cM of the lead SNP for each disease and region were selected. The Sept 2009 release comprising 60 CEU individuals was used (~5x mean read-depth whole genome sequencing) selecting markers called in at least two of the Broad Institute / Sanger Institute / University of Michigan algorithms. Additional genomic region re-sequencing content was submitted for Immunochip at specific loci from coeliac disease, Crohn's disease and type 1 diabetes cases and controls (Supplementary Note).

Genotyping. Samples were genotyped using the Immunochip as per Illumina's protocols (at labs in London, Hinxton, Groningen and Charlottesville). NCBI build 36 (hg18) mapping was used (Illumina manifest file Immuno_BeadChip_11419691_B.bpm).

Data Quality Control. Very low call rate samples and variants were first excluded (and samples repeated). The Illumina GenomeStudio GenTrain2.0 algorithm was used to cluster an initial 2,000 UK samples. Subsequently with additional sample data (case and control data were analysed together) clusters were re-adjusted or excluded (manual or automated) for variants with low quality statistics (call rate<99.5%, low GenCall score, many high-intensity no-calls). This method was superior to the GenoSNP or Illuminus clustering algorithms (not shown). A cluster set based on 172,242 autosome/X-chromosome variants (available on request) was then applied to all samples. Samples were excluded for call rate <99.5% across 172,242 markers. We then removed 15,657 non-polymorphic markers (i.e. only one of three expected genotype clouds observed) which reflect a combination of ethnic-specific variants, allele-specific assay failure, as well as substantial false-positive rates in early next-generation sequencing SNP calling algorithms.

Samples were excluded for incompatible recorded gender and genotype inferred gender, duplicates and first/second degree relatives. Potential ethnic outliers were identified by multi-dimensional scaling plots of samples merged with HapMap3 data, the subset of SNPs common to HapMap3 and Immunochip accurately identified the different HapMap3 population samples. We considered the white European and Indian collections separately.

Stepwise conditional logistic regression is sensitive to missing data and subtle genotyping error, and we therefore desired an ultra-high quality dataset. Markers were excluded from all sample collections for deviation from Hardy-Weinberg equilibrium in controls (P<0.0001) and/or differential missingness in genotype no-calls between cases and controls (P<0.001), in any of the seven collections. Finally we required a per-SNP call rate of >99.95% (a maximum of

12 no-call genotypes from 24,269 samples per autosomal marker), generating a dataset of 139,553 markers (of which all but 372 indels are SNPs).

We visually inspected intensity plot genotype clouds for all markers described in Table 2 (and further potential loci with $P<1.9x10^{-6}$), and confirmed all to be high quality.

Genotype data has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00000000053.

Statistical Analysis. Analyses were performed using PLINK v1.07[38], using logistic regression tests with gender as a covariate, and collection membership (Table 1) as a factorized covariate. Stepwise conditional logistic regression was performed, in order of markers with the smallest P value. Graphs were plotted in R, and using a modified version of LocusZoom[39].

We permuted affection status for the dataset at each region (Online Methods), to establish locus wide statistical significance thresholds for defining independently associated SNPs. For each locus, given by the linkage disequilibrium boundaries (Table 2) we calculated the 5th percentile based on the nominal *P*-value distribution for 1000 permutations controlling for multiple marker testing. This approach proved slightly more stringent than a per-locus Bonferroni correction for independent (using an estimate for "independence" of pairwise $r^2<0.05$) variants (Supplementary Table 3). We estimated our dataset contained 26,146 completely uncorrelated variants (using pairwise $r^2<0.05$ and a sliding 1000 SNP window).

The fraction of additive variance was calculated using a liability threshold model[40] assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the UK dataset. Genetic variance was calculated assuming 50% heritability.

ACKNOWLEDGMENTS

We thank Coeliac UK for assistance with direct recruitment of celiac disease individuals, and UK clinicians (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis, K. Moriarty) who recruited celiac disease blood samples described in our previous studies. We thank the Dutch clinicians for recruiting celiac disease blood samples described in our previous studies (C.J. Mulder, G.J. Tack, W.H.M. Verbeek, R.H.J. Houwen, J.J. Schweizer). We thank the genotyping facility of the UMCG (Pieter van der Vlies) for helping in generating part of immunochip data and S. Jankipersadsing, A. Maatman, at UMCG for preparation of samples. We thank R. Scott for preparing samples for genotyping and the University of Pittsburgh Genomics and Proteomics Core Laboratories for performing genotyping. We thank C. Wallace for assistance with Immunochip SNP selection, and J. Stone for co-ordinating Immunochip design and production at Illumina. We thank the members of each disease consortium who initiated and sustained the cross-disease Immunochip project. We especially thank all individuals with celiac disease and control individuals for participating in this study.

Funding was provided by the Wellcome Trust (084743 to D.A.vH.); by grants from the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative, and partially funded by the Dutch Government (BSIK03009 to C.W.) and the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.); by NIH grant 1R01CA141743 (to R.H.D); Fondo de Investigación Sanitaria FIS08/1676 and FIS07/0353 (to E.U.). This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, the National Human Genome Research Institute, the National Institute of Child Health and Human Development, and the Juvenile Diabetes Research Foundation International and is supported by NIH grant U01-DK062418. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS-CC collection), funded by the Wellcome Trust grant 076113/C/04/Z and by NIHR programme grant to NHSBT (RP-PG-0310-1002). The collection was established as part of the Wellcome Trust Case Control Consortium (WTCCC)[33]. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

AUTHOR CONTRIBUTIONS

DAvH and C. Wijmenga led the study. Major contributions were (i) DAvH, KAH, GT and C. Wijmenga wrote the paper; (ii) KAH, GT, VM, NB, JR, MP, MM, RHD and KF performed DNA sample preparation and genotyping assays; (iii) DAvH, VP, KAH, GT performed statistical analysis. Other authors contributed mainly to sample collection and phenotyping. PD led the formation of the Immunochip Consortium, with SNP selection by JB and C. Wallace. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1.Bingley, P.J. *et al.* Undiagnosed coeliac disease at age seven: population based prospective birth cohort study. *BMJ* **328**, 322-3 (2004).

2.West, J. *et al.* Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. *Gut* **52**, 960-5 (2003).

3.van Heel, D.A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat Genet* **39**, 827-9 (2007).

4.Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* **40**, 395-402 (2008).

5.Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**, 295-302 (2010).

6.Trynka, G. *et al.* Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. *Gut* **58**, 1078-83 (2009).

7.Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature reviews. Genetics* **10**, 43-55 (2009).

8.Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N Engl J Med* **359**, 2767-77 (2008).

9.Zhernakova, A. *et al.* Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. *PLoS genetics* **7**, e1002004 (2011).

10.Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis research & therapy* **13**, 101 (2011).

11.Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

12.Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature genetics* **37**, 1243-6 (2005).

13.Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature genetics* **42**, 1049-51 (2010).

14.Spencer, C., Hechter, E., Vukcevic, D. & Donnelly, P. Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS genetics* **7**, e1001337 (2011).

15.Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-8 (2010).

16.van Heel, D.A., Hunt, K., Greco, L. & Wijmenga, C. Genetics in coeliac disease. *Best Pract Res Clin Gastroenterol* **19**, 323-39 (2005).

17.Zhernakova, A. *et al.* Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am J Hum Genet* **86**, 970-7 (2010).

18.Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature genetics* (2011).

19.Lesage, S. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *American journal of human genetics* **70**, 845-57 (2002).

20.Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics* **42**, 684-7 (2010).

21.Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annual review of genetics* **44**, 293-308 (2010).

22.Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS biology* **8**, e1000294 (2010).

23.Zheng, Q. & Wang, X.J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic acids research* **36**, W358-63 (2008).

24.Lanktree, M.B. *et al.* Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. *American journal of human genetics* **88**, 6-18 (2011).

25.Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728-31 (2008).

26.Lowe, C.E. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature genetics* **39**, 1074-82 (2007).

27.Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841-5 (2010).

28.Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nature genetics* **43**, 801-5 (2011).

29.Jostins, L., Morley, K.I. & Barrett, J.C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *European journal of human genetics : EJHG* **19**, 662-6 (2011).

30.Asano, A., Tsubomatsu, K., Jung, C.G., Sasaki, N. & Agui, T. A deletion mutation of the protein tyrosine phosphatase kappa (Ptprk) gene is responsible for T-helper immunodeficiency (thid) in the LEC rat. *Mammalian genome : official journal of the International Mammalian Genome Society* **18**, 779-86 (2007).

31.Adrianto, I. *et al.* Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. *Nature genetics* **43**, 253-8 (2011).

32.Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-9 (2010).

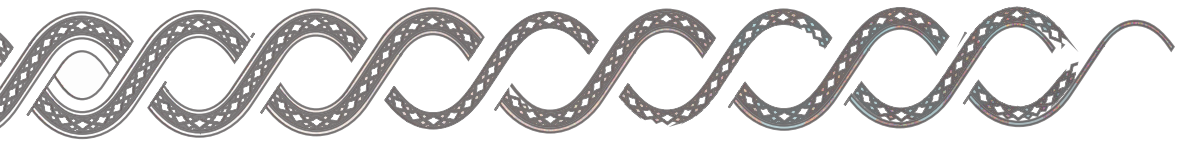
33. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).

34. Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition. *Archives of disease in childhood* **65**, 909-11 (1990).

35.Romanos, J. *et al.* Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. *Journal of medical genetics* **46**, 60-3 (2009).

36.Plaza-Izurieta, L. *et al.* Revisiting genome wide association studies (GWAS) in coeliac disease: replication study in Spanish population and expression analysis of

candidate genes. *Journal of medical genetics* **48,** 493-6 (2011).
37.Megiorni, F. *et al.* HLA-DQ and risk gradient for celiac disease. *Human immunology* **70,** 55-9 (2009).
38.Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81,** 559-75 (2007).
39.Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336-7 (2010).
40.Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405,** 847-56 (2000).

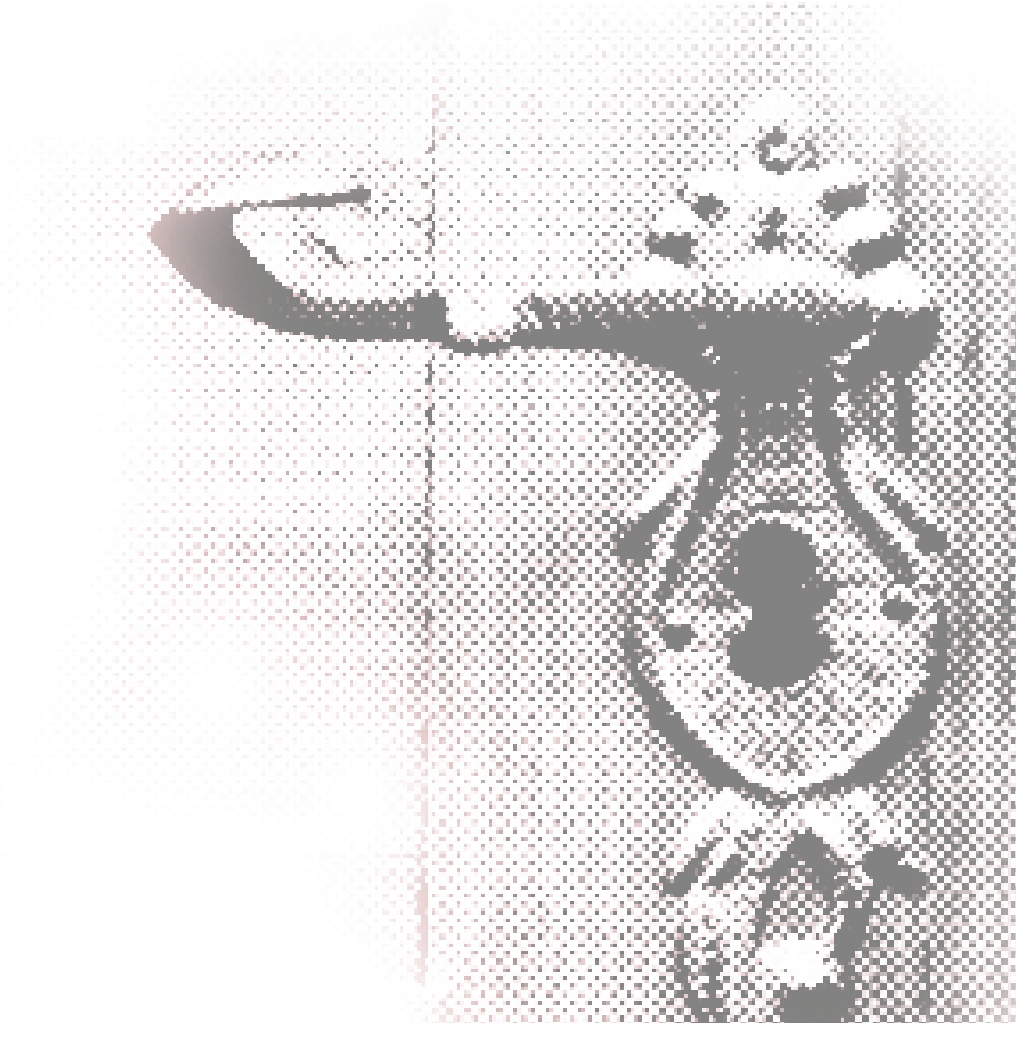


Dense genotyping replicates and further localizes coeliac disease association signals in a north Indian population

Submitted

Gosia Trynka*, Sabyasachi Senapati*, Jihane Romanos, Agata Szperl, Alexandra Zhernakova, Ajit Sood, Vandana Midha, Lude Franke, Santos Alonso, Thelma BK, Cisca Wijmenga

*equal contribution



ABSTRACT

Genome-wide association studies (GWAS) in coeliac disease have shown strong and replicable association to 26 non-HLA loci in individuals of European descent. Since coeliac disease is equally prevalent in northern India, we aimed to assess association to these 26 risk loci in a north Indian population and to see whether dense SNP genotyping in both populations could help refine the association signals. We first performed cross-ethnic mapping in 497 coeliac disease patients and 736 controls from northern India and in 1150 Dutch cases and 1173 controls; we used Immunochip, a customized genotyping platform providing a higher SNP resolution than standard GWAS platforms. Due to differences in linkage disequilibrium (LD) structure between Europeans and north Indians, we applied two different strategies, which led to replication of 50% of the non-HLA loci. In an 'exact' analysis – directly testing for association of the index SNP – we replicated three loci, but in a 'transferability' analysis – testing the exact European SNP along with the variants in LD – we replicated 12 loci. The use of the densely probed Immunochip increased the replication rate by 15% compared to the standard GWAS platform. We also showed that four of the 26 non-HLA loci were under positive selection pressure (*IL18RAP, CCR1/5, SH2B3* and *PLEK*). Secondly, we narrowed down the association signals by using a cross-population analysis. We observed that the northern Indian population was characterized by less extensive long-range LD than our Dutch population. For five of the 26 non-HLA loci, this resulted in finer localization of the association signal and for another five loci we observed a shift in the association signal. Remarkably, the long-range LD was broken down at the 592 kb *IL2/IL21* locus and localized the association signal to a very small interval of 21 kb, near the promoter region of the *KIAA1109* gene.

INTRODUCTION

Over the past five years, genome-wide association studies (GWAS) have had tremendous success in mapping genetic loci for common, complex diseases. To date, more than 4000 SNPs have been reported to be associated with the risk for more than 400 distinct traits[1]. Combining immune-mediated diseases has already established some 180 common variants with modest effect sizes and many more variants are expected to be identified that account for part of the remaining "hidden heritability". The great majority of GWAS for immune-related diseases have been conducted in populations of European ancestry and it is likely that further risk alleles may be identified in populations with different ethnic backgrounds[2]. However, our knowledge of the genetic architecture of common complex diseases across multi-ethnic cohorts is rather limited and there is a clear need to replicate disease-associations across different ethnicities. A limited number of recent GWAS and replication studies in multi-ethnic cohorts, including African or Asian populations, suggest that Caucasian association signals can be generalized to populations with other ethnic backgrounds[3-5]. On the other hand, a literature search also showed that 25-55% of the association signals at shared loci are independent between populations (distance > 500 kb in each population). This suggests that disease aetiology is common between populations but that risk variants are often population-specific[6]. This observation has consequences for the design of cross-ethnicity replication studies.

Replication studies in ethnically different populations often attempt to replicate only the reported, and therefore most significant, SNP from a certain study or combined meta-analysis. Given the difference in linkage disequilibrium (LD) structure between distinct ethnic groups, such an approach may fail if this top-SNP 'poorly' tags the true risk variant in another population. Alternatively, the most significant and reported SNP, together with SNPs in LD with it, can be tested in additional and ethnically different samples[7]. One obstacle is, however, that the available GWAS platforms have been designed to capture genetic variation in populations of Caucasian descent and have less power for other ethnic groups. We aimed to overcome some of these limitations by performing a cross-ethnic study of established risk loci in a northern Indian case-control cohort.

Coeliac disease is an autoimmune inflammatory disease of the small intestine, caused by interaction with gluten in genetically predisposed individuals. It is the most common form of intestinal inflammatory disorder among Europeans with 1-3% prevalence[8], and a similar 1% prevalence in the Indian sub-continent[9]. The largest genetic risk to coeliac disease is conferred by variants in HLA genes, which account for approximately 35-40% of the genetic risk[10]. Recent progress in understanding the genetic architecture of coeliac disease has led to the identification of 26 non-HLA loci in a meta-analysis of British, Dutch, Italian and Finnish populations, with replication in further cohorts of European origin[10].

In this study, we focused on the 26 established non-HLA coeliac disease loci and assessed their replication in a north Indian population-based sample of 497 patients and 736 unrelated controls. We combined our replication study with fine-mapping, as we used information on 15,851 SNPs across these 26 loci (i.e. on average 610 SNPs per locus) which were present on the Immunochip. The Immunochip is a custom-made genotyping platform with ~200,000 markers, of which the great majority map within the 183 loci associated to immune-related diseases[11,12]. This platform was specifically designed to fine-map currently known GWAS loci and results, on average, in a 12-15x greater marker density than a standard GWAS chip.

We applied two different strategies to assess the genetic architecture of coeliac disease in the north Indian population: (1) an 'exact' analysis, directly testing the index top-associated SNP reported for Europeans, and

Locus	Index SNP	Top Transferable SNP	Associated allele	BP	MAF		P-value		Odds ratio		r^2/D' with index SNP	
					Dutch	North Indian	Dutch	North Indian	Dutch	North Indian	CEU (1000 Genomes)	north Indian
PLEK	rs17035378 [imm_2_68452459]	rs9309419 [imm_2_68521802]	A	68521802	0.18	0.18	0.043	0.00110	1.16 [1.00-1.34]	0.66 [0.52-0.85]	0.05 0.68	0.03 0.32
IL18RAP	rs7559479 [imm_2_102435219] $^{\delta}$	rs4851005 [imm_2_102377984]	T	102377984	0.39	0.32	0.069	0.00900	0.87 [0.77-0.98]	0.77 [0.64-94]	0.16 1.00	0.33 0.93
ITGA4/ UBE2E3	rs1018326 [imm_2_181716045] $^{\delta}$	NA [imm_2_181849926]	G	181849926	0.36	0.32	0.627	0.00130	0.97 [0.85-1.09]	0.68 [0.54-0.85]	0.21 0.78	0.25 0.62
CCR1/5	rs13098911 [imm_3_46210205]	rs17999888 [imm_3_46387263]	C	46387263	0.46	0.42	0.041	0.00360	1.12 [1.00-1.26]	0.76 [0.64-0.92]]	0.06 0.69	0.03 0.37
IL12A	rs17810546 [imm_3_161147744]	rs1498736 [imm_3_161177252]	A	161177252	0.22	0.39	0.007	0.00293	0.82 [0.72-0.96]	0.75 [0.62-0.91]	0.05 1.00	0.04 1.00
IL2/ IL21	rs13151961 [imm_4_123334952]	rs6534338 [imm_4_123246319]	T	123246319	0.32	0.16	0.004	0.00141	1.20 [1.06-1.36]	1.44 [1.13-1.62]	0.08 1.00	0.03 1.00
THEMIS/ PTPRK	rs802734 [imm_6_128320491]	rs4142030 [imm_6_128196379]	G	128196379	0.38	0.4	0.039	0.00088	1.14 [1.01-1.27]	1.36 [1.24-1.90]	0.22 0.59	0.12 0.47
TNFAIP3	rs2327832 [imm_6_138014761]	NA [imm_6_137980782]	A	137980782	0.49	0.33	0.333	0.00952	0.95 [0.85-1.07]	1.28 [1.06-1.54]	0.08 0.72	0.12 0.59
TAGAP	rs1738074 [imm_6_159385965]	rs9347286 [imm_6_159435348]	T	159435348	0.15	0.08	0.083	0.00167	0.86 [0.73-1.02]	0.55 [0.38-0.80]	0.1 0.79	0.002 0.20
ZMIZ1	rs1250552 [imm_10_80728033]	rs1250549 [imm_10_80730480]	T	80730480	0.47	0.44	0.0005	0.00814	0.80 [0.72-0.90]	1.27 [1.06-1.51]	0.78 0.90	0.75 0.93
ETS1	rs11221332 [imm_11_127886184]	NA [imm_11_127904148]	T	127904148	0.22	0.21	0.242	0.00296	0.90 [0.78-1.04]	0.70 [0.56-0.89]	0.09 1.00	0.08 0.93
ICOSLG	rs2838531 [imm_21_44463014] $^{\delta}$	rs6518350 [imm_21_44446245]	G	44446245	0.19	0.16	0.178	0.00075	0.89 [0.77-1.03]	0.62 [0.47-0.82]	0.48 0.91	0.27 0.67

Table 1. Summary statistics of the 12 transferable loci. Published European SNPs that were not present either on Immunochip or in the 1000 Genomes data were replaced by their perfect proxies (δ).

(2) a 'transferability' analysis[7], testing the exact European SNP along with the variants in LD (r^2 >0.05) which were present on the Immunochip. Our main objectives were to evaluate the known coeliac disease-associated loci in an ethnically distinct north Indian population, and to refine the association signals. Using this approach, we convincingly replicated 13 of the 26 loci (50%). Had we tested only the reported European variant using the 'exact' approach, the replication success rate would have been 19%. We further performed the cross-ethnic fine-mapping by comparing LD structures between the north Indian and Dutch cohorts for all 12 transferable loci and assessed the signatures of selection pressure acting upon them.

RESULTS

We obtained high quality genotype data for 160,448 polymorphic variants that were common in 497 north Indian coeliac disease cases and 736 north Indian controls, and in 1150 Dutch coeliac disease cases and 1173 Dutch controls. The first three components of multi-dimensional scaling analysis (MDS) showed that the north Indian samples overlapped with the GIH samples from HapMap3 (Gujarati Indians in Houston, Texas) (Figure_S1). However, the analysis of GIH and our north Indian samples indicated subtle population substructures. The second component separated the north Indian samples from the GIH samples, while the third component identified further gradients in both these populations. The fact that the first and third components seemed to extract similar information from both the GIH and our cohort indicates that the bias is not due to our cohort sampling but is a genetic characteristic of this ethnic population.

After stringent quality control (see Material and Methods), we obtained 17,979 SNPs mapping within the 26 non-HLA loci, 15,851 of which were polymorphic in at least the north Indian or the Dutch cohort and therefore informative for further analysis. Of the set of 15,851 SNPs, 8,664 (54.7%) were variants with a low minor allele frequency (MAF<0.05) in the north Indians and 8,484 (53.5%) in the Dutch.

Replicating European signals in north Indian cohort

As expected, we observed strong association of the north Indian coeliac disease cases to the HLA region (rs2854275, $p=2.634^{-49}$). We observed strong inflation in the test statistics among all Immunochip variants (λ=2.91), which is mainly driven by the HLA locus (Figure 1A) as exclusion of the HLA region SNPs reduced the λ to 1.04 (Figure 1B). As the Immunochip is preselected for "immune" SNPs, it is difficult to distinguish enrichment of true signals from population stratification. When we

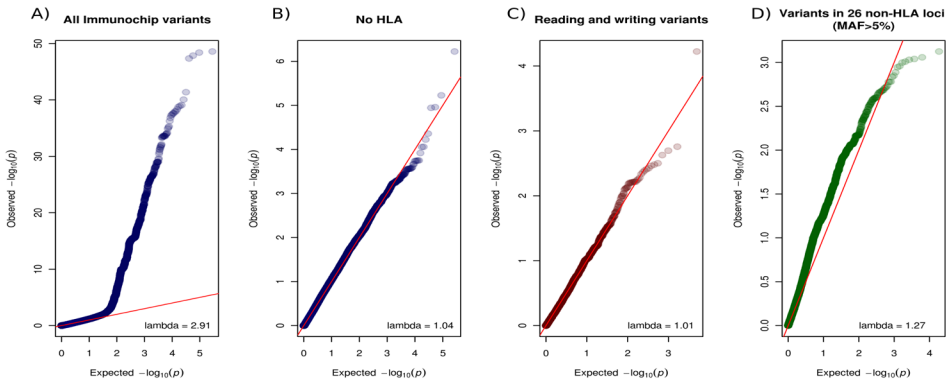


Figure 1 We observed strong inflation in the test statistics for all the Immunochip variants when including the HLA region (λ=2.91; panel A). This inflation decreased to λ=1.04 after excluding some 10,000 markers from the HLA locus, a well known, strong genetic risk factor for coeliac disease (panel B). The SNPs submitted for replication of the 'reading and math skills' were used as 'null' variants not confounded by the immune signal, to test for population stratification in our north Indian cohort (λ of 1.01; panel C). Variants in the 26 non-HLA, coeliac disease loci were strongly enriched for association signals (panel D).

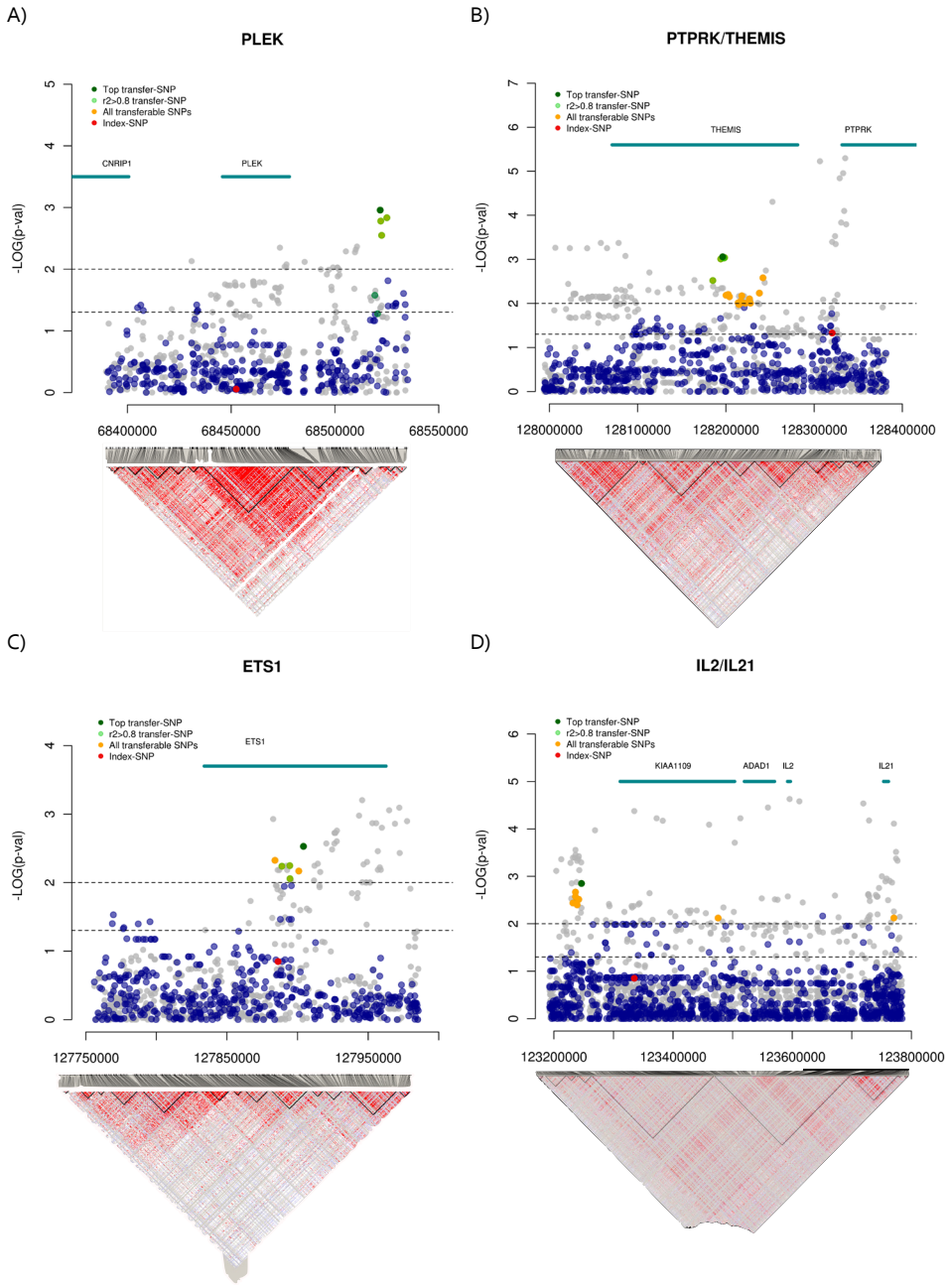


Figure 2 Examples of cross-ethnic mapping results. A) Shifted and more localized association signal in the north Indian samples compared to the Dutch association. Note that the association signal in the smaller north Indian samples is stronger than in the Dutch samples, suggesting that some loci may have stronger effects in one population than the other. PLEK is also a locus for which we observed positive selection pressure. B) A cluster of correlated variants localizes in the intronic region of THEMIS gene, whereas the Dutch signal maps 124 kb upstream, pointing towards the 3'UTR of the PTPRK gene. At this locus the index SNP was successfully replicated in north Indians ($P<0.05$). C) Overlapping signal between the Dutch and north Indians, with tighter SNP clustering over a 21 kb block in north Indians, compared to the dispersed association over a 103 kb block in the Dutch. D) The overlapping signal between the two populations is broadly spread across the region. However, in north Indians a cluster of correlated SNPs localizes in the small 21 kb LD block. Note the LD breakage in north Indians due to the larger number of low frequency SNPs (MAF<0.1), which results in two smaller LD blocks.
Dark blue represents the association signal in the north Indian samples and grey in the Dutch. Dark green depicts the most associated transferable SNP, and light green the SNPs strongly correlated with it (r2>0.08). Yellow depicts all the transferable SNPs and red indicates the index SNP. Colours may overlap.

used the 2,626 'null variants' present on the Immunochip (see Materials and Methods), we noted that there was no inflation (λ=1.01), implying that the observed associations were not confounded by population substructure (Figure 1C). We observed a clear indication of association for the 15,851 SNPs residing in the 26 non-HLA coeliac disease loci (λ=1.27, Figure 1D).

The initial association of coeliac disease to the 26 non-HLA loci, all of which have modest effects on the genetic risk (OR of 0.74 -1.36), was performed in a sample size of over 5,000 coeliac cases and 10,000 controls[10]. We realize that our north Indian cohort was underpowered to detect associations with small effects although we did reach 80% power for common SNPs (MAF>0.25) and OR>1.3 (Figure S2).

'Exact' analysis

For the 'exact' analysis we aimed to test the same top-associated SNP reported for Europeans ('index' SNP), as described by Dubois et al.[10], in our north Indian cohort. All of the 26 European index SNPs were polymorphic and frequent (MAF>0.05) in north Indians, although the correlation of the minor allele frequencies for these 26 index SNPs between the north Indians and Dutch was much lower (r^2=0.35) than the correlation between European cohorts (mean r^2= 0.9) (Figure S3). Of the 26 index SNPs, five were directly replicated at p< 0.05, with the same direction as in Europeans (Table S1). The five SNPs tagged the following genes: *IL12A* (rs17810546, p = 0.02), *ICOSLG* (rs4819388, p = 0.035), *ZMIZ1* (rs1250552, p =0.04), *THEMIS/PTPRK* (rs802734, p=0.047), and *LPP* (rs1464510, p=0.048) loci. Of the remaining 21 variants that did not reach the replication significance threshold, 14 (66%) showed the same directionality as in Europeans, which is more than the 50% chance expected under the null hypothesis (p=0.05, binomial probability, we excluded two loci for which OR was 1) (Figure S4).

'Transferability' analysis

We next performed an analysis that accounts for the different LD patterns between Europeans and north Indians. In this we tested all the variants in each of the 26 non-HLA loci that were in LD with the top SNP (r^2 > 0.05, based on CEU, 1000 Genomes Project). To claim a locus as 'transferable', at least one SNP per locus had to be associated at p<0.01 in our north Indian sample. Twelve of the 26 loci were transferable to Indians, which was significantly more than expected by chance ($p_{permuted}$ = 0.031). These were: *PLEK, IL18RAP, ITGA4/UBE2E3, CCR1/5, IL12A, IL/IL21, THEMIS/PTPRK, TNFAIP3, TAGAP, ZMIZ1, ETS1* and *ICOSLG* (Table 1,

Author summary

Most disease gene mapping approaches have been conducted in populations of European descent. It is unclear how many of these findings can be translated across different racial/ethnic groups: are the risk alleles the same, are the effects of associated alleles comparable, and how comparable are the localization signals? Fine-mapping is one approach to better localize the signals and thereby bring us closer to translational studies and a practical application of genetic findings. Cross-ethnic mapping in distinct populations has the potential to aid such fine-mapping efforts. We applied this mapping strategy to fine-map 26 non-HLA loci associated to coeliac disease in Europeans. For the first time, we used dense genotyping information from a custom-made platform to elucidate the best possible linkage disequilibrium resolution. We replicated 50% of these risk loci in a northern Indian cohort. For five of the regions, the signals from the north Indian population clustered separately from those in Europeans, which allowed us to refine the association signal. In five other loci, we were able to narrow down the association to a smaller genetic interval, covering a cluster of correlated markers. This indicates that we may have identified regions which are likely to capture the causative variants.

and Table S2 for details on non-transferred loci); the *IL12A, THEMIS/PTPRK, ZMIZ1* and *ICOSLG* loci were also identified by the 'exact' analysis. None of the transferable variants demonstrated significant cross-population heterogeneity between north Indians and the Dutch (Breslow-Day $p_{permuted}<1.27x10^{-5}$).

Coverage and transferability rate

To investigate how much we gained by using a dense genotyping platform in contrast to a standard GWAS chip, we calculated the transferability success rate of the 1,284 SNPs that map within the 26 coeliac disease loci that are shared between the Illumina Hap550 platform and Immunochip. We replicated nine loci at p<0.01, yielding a 35% success rate compared to the 50% when using the Immunochip (Table S3).

The transferability rate will be strongly influenced by the actual coverage of the genetic variation. We therefore assessed how well the Immunochip covers total variation at coeliac disease loci in the 1000 Genomes European samples. There are 26,863 polymorphic variants annotated within coeliac disease loci in CEU, TSI, FIN and GBR individuals, 12,497 of which (46%) are shared with Immunochip SNPs (821 were non-polymorphic in either the Indians or Dutch). 648 SNPs were genotyped on Immunochip but were not present in the 1000 Genomes data. 5,632 variants (39%) of the non-genotyped 1000 Genomes variants (14,366) contribute to the common variation in Europeans (MAF>0.1). The remaining variation is of low frequency (MAF<0.05, Figure S5), and it is likely that these are largely population-specific or falsely called sequence variants.

Cross-population LD measure

To assess if the north Indian transferability signals reflect European associations, we estimated the degree of LD correlation (quantified by pair-wise SNP LD, r^2) between the European index SNP and the north Indian top transferable SNP in CEU and north Indian data. We observed that the north Indian transferable signals reflected the European LD (correlation coefficient r^2=0.83) (Figure S6A). However, the majority of the transferable signals accumulated in the lower tail of r^2 values, with a small shift towards less tight LD in north Indians than in the CEU population.

We also evaluated the LD correlation between all transferable variants (all SNPs in r^2> 0.05 based on CEU, 1000 Genomes data, and present in north Indians) and the top SNP (Figure S6B). We observed similar results, although the LD correlation between CEU and Indians was lower (r^2=0.55), most of the transferable variants were located in the lower tail of r^2 values.

Contrasting association signals

As a consequence of longer LD blocks, associations in Europeans often map to regions containing multiple genes and because of the strong LD, it can be difficult to pinpoint the true risk gene. Using ethnically distinct populations offers the opportunity to fine-map signals because of population differences in LD structure.

From the 13 loci that were replicated in north Indians, we excluded the LPP locus because it was only replicated by the direct SNP test. We then compared the overlap of the association signal patterns between north Indians and the Dutch and evaluated the LD structure (Figure S7). For five loci (*IL18RAP, CCR1/5, IL12A, IL2/IL21* and *ETS1*), the pattern of association in north Indians was consistent with that seen in the Dutch. Nonetheless, for some loci, the association signal could indeed be refined to a smaller region. Although in the 2q12.1 region (*IL18RAP* locus) the European and north Indian signals overlap, the LD is lower in the north Indians and the main LD block covering the association signals is 7 kb smaller than in the Dutch. This LD block covers *IL18R1, IL18RAP* and part of *SLC9A4* but clearly excludes *IL1RL2* and *IL1RL1* (Figure S7). The only transferable SNP mapped in the intron of *IL18R1*. Both Dutch and north Indian association signals show very high D' with the European Index marker (Table 1). At the *CCR1/5* locus, both populations show a similar LD background and

overlapping association signals, which narrows down the association signal to a smaller region of ~250 kb, including the *CCR1, CCR2, CCR3, CCR5, CCRL2* and *LTF* genes. The north Indian top transferable SNP mapped in the exon 2 of *CCR5*.

At the *IL12A* locus, association signals localized in the intergenic region between *SCHIP1* and *IL12A* with the top transferable north Indian SNP mapping in a small LD block of 6 kb (161176264 bp – 161182066 bp) near the promoter of *IL12A*. The coeliac disease region at chromosome 4q27 harbours four genes, including two plausible candidates for coeliac disease, *IL2* and *IL21*. Due to the strong LD across the locus, the signal cannot be further refined. However, in our north Indian cohort the LD was broken down due to the larger number of low frequency SNPs (MAF < 0.1) (Figure 2D, Figure S7), resulting in two smaller LD blocks in north Indians, compared with one large block in the Dutch (Figure S8D). The association signal was also spread along the whole region, although an outstanding cluster of most strongly associated SNPs was located in the small 21 kb LD block (123246379 bp - 123267309 bp) adjacent to the large LD block that covers the *KIAA1109* gene and top European SNP.

At the *ETS1* locus, the LD architecture is very similar between the Dutch and north Indians and the cross-population signals overlap (Figure 2C, Figure S7). However, the north Indian signal is much more tightly clustered around the top European SNP (21 kb block, from 127882690 bp to 127904148 bp), whereas the Dutch signal is widely spread (a 103 kb region from 127882690 bp to 127985506 bp).

At the *TNFAIP3* and *ZMIZ1* loci only a single SNP was transferable, hence it was not possible to deduce the association patterns. At *ZMIZ1* locus, the top transferable north Indian SNP is localized in proximity to the European index SNP in the intronic part of the *ZMIZ1* gene (Figure S7).

The remaining five loci showed a shift in the association pattern. At 2p14 the Dutch signal spreads over the middle LD block and covers the *PLEK* gene, whereas the north Indian signal locates at 69 kb from the European top SNP, downstream of this gene (Figure 2A, Figure S7). The north Indian signal is located in a block of low LD and is poorly correlated with the European index signal. At 2q31.3 the north Indian association signal is stronger than that in the Dutch population. In the Dutch, this region is covered by three LD blocks, with the association signal mapping in the second block, an intergenic region downstream of the *UBE2E3* gene. In north Indians the LD is even more broken down and the association signal localizes in two clusters of SNPs: first, in a small block in close proximity to the top European SNP, and second, in a distal and stronger signal in the block partly corresponding to third LD region. (Figure S7). At the 6q22.33 locus we observed a cluster of correlated variants in the intronic region of the *THEMIS* gene. Interestingly, however, the Dutch signal at this locus maps 124 kb upstream, clearly pointing towards the 3'UTR of the *PTPRK* gene (Figure 2, Figure S7). At the 6q25.3 locus, the European top SNP is located in the first exon of the *TAGAP* gene, whereas the north Indian signal is located in the promoter region, which partly also overlaps with the Dutch signal (Figure S7). At the *ICOSLG* locus, the signal overlapped with the European one to some degree, but we noticed a tight cluster of SNPs further upstream of the *ICOSLG* gene (Figure S7), suggesting that the causal variant could be captured within the 17 kb LD block (from 44435321 bp to 44452009 bp).

Evaluation of the selection pressure

To understand evolutionary variability at the 26 coeliac disease loci, we estimated a pair-wise fixation index (Fst) using the north Indian and Dutch cohorts. We tested Fst for all 12 reported index SNPs (Table S4). We found suggestive evidence for positive selection pressure for previously reported *IL18RAP* (rs7559479), *CCR1/5* (rs13098911), *PLEK* (rs17035378) and *SH2B3* (rs653178) loci (Fst>0.079) values.

DISCUSSION

Recent cross-ethnic studies indicate that a large proportion of disease- or trait-associated common variants established in populations of European descent also contribute to the risk in other ethnic groups[3,4,13]. Yet this observation cannot be generalized because only a limited number of the GWAS were conducted in non-European populations[14]. In this study we sought to replicate and narrow down coeliac disease loci associated in Europeans in a north Indian cohort. To efficiently capture the differences in genetic background between two ethnically distinct populations we used the Immunochip, a genotyping platform that, at present, offers the highest coverage of regions associated to immune-related disorders. We directly tested for association of the top European SNP (exact test) and also performed association of the 'index' SNP together with variants in LD with it (transferability test). When investigating the LD correlation between index SNPs and transferable markers we noticed a trend towards a lower range of r^2 values in north Indians than in the Dutch (Figure S6). This indicates a lesser degree of tagging properties between the index SNP and transferable markers in north Indians, which could negatively influence our replication success rate. Of the 26 coeliac disease loci established in Europeans, five were replicated by the exact index SNP association analysis and 12 were replicated by the transferability approach, yielding a total of 13 unique, replicated loci. Overall, our replication success rate was 50%, compared to the 19% that we would have obtained by testing only the European index SNP.

We observed an advantage of using the Immunochip over a standard GWAS platform, with a 50% vs. 35% successful transferability rate. Nonetheless, despite the high marker density, we estimate that the Immunochip only covers approximately 50% of the European genetic variants (based on the 1000 Genomes data, release May 2011) and mainly misses the low frequency variants. Low and rare frequency variants are more likely to be population-specific[14], therefore even if they were included on the Immunochip they might have been of limited value. At the moment there is no sequence data available for the north Indian populations and it is difficult to estimate how well the Immunochip performs in this population. Most likely our transferability success rate is underestimated, while differences in haplotype frequencies and linkage disequilibrium mean it is likely that the Immunochip covers the genetic variation in north Indians less well than in Europeans.

India has been underrepresented in genome-wide associations and our study is the first to perform an association study on coeliac disease in north Indians and to establish the association of 13 loci in this population. The lack of replication of the remaining 13 European loci may be due to the limited power of our study, and/or poor tagging properties of the causal variant of the tested SNPs in north Indians, but it is also likely that some European loci will not confer risk for coeliac disease in north Indians. Interestingly, despite the fact that our north Indian cohort was half the size of the Dutch cohort, the association signals at the *ITGA4/UBE2E3, PLEK, TAGAP* and *ICOSLG* loci were equal or stronger in the north Indians than those in the Dutch. This indicates that some regions may confer different risk burdens to different populations.

North Indians, although genetically close to Europeans, differ markedly in their allele frequencies from Europeans, which is partly reflected by a lesser degree of LD structure compared to Europeans[15]. The different allele frequencies and lower extent of long-range LD in north Indians resulted in different association signal localization, or allowed finer-scale mapping at some loci. For example, the *IL2/IL21* locus consists of a large LD block of 315 kb that covers the majority of the 591 kb region with highly correlated markers. It contains four genes: two are interleukins, *IL2* and *IL21*, one plays a role in spermatogenesis (*ADAD1*), and one is a transcript of unknown function (*KIAA1109*). Because of its very strong LD, this region has not been fine-mapped successfully

using European samples. Our study shows that north Indians have a higher proportion of low frequency markers (MAF<0.1) than the Dutch, which breaks down the LD structure. The association signal localizes in the small 21 kb LD block (from 123246319 bp to 123267319 bp) adjacent to the larger block covering the *KIAA1109* gene and the top European associated SNP. Although we narrowed down the association signal to the 21 kb block, it does not include a gene, which could suggest that the causal variant plays a regulatory role, and although the associated LD block is in the immediate proximity of *KIAA1109*, the variant may also impact the more distal interleukin genes. This possibility, however, needs to be followed up, preferably by sequencing the region in the north Indians to identify all variants in LD with the top transferable SNP. Thus, although we were not able to point to the likely causal gene in this region, we did succeed in narrowing down the association to a small genetic interval. Another successful example is the *ETS1* locus, where the north Indian signal is limited to tightly clustered variants, suggesting that the causative variant resides within a small 21 kb genetic interval (from 127882690 bp to 127904148 bp).

Differently localized association signals could also indicate locus heterogeneity and different biological pathways underlying the disease aetiology. In our previous GWAS (10), in which we established the association at the 6q22.33 locus, we suggested *THEMIS* as a plausible candidate gene. However, fine-mapping in the Europeans using Immunochip clearly points towards the neighbouring *PTPRK* gene, a protein tyrosine phosphatase, whereas the north Indian signal points towards *THEMIS*. Unless functional follow-up studies are conducted to confirm a gene's causality, it is unclear which gene truly plays a role in the disease pathogenesis. Nevertheless, it is tempting to speculate that *PTPRK*, as a signalling molecule regulating a variety of cellular processes including cell growth or differentiation, is a less attractive gene candidate for an immune disease than the neighbouring *THEMIS* gene, which regulates positive and negative T-cell selection during thymocyte development. *THEMIS* is transcribed from the reverse strand and, in fact, both associations could still point to the same gene, for example in the Europeans by affecting the transcription regulation and in the north Indians by affecting a regulatory mechanism within the gene or by altering the gene structure. This requires follow-up studies for further elucidation.

We found suggestive evidence of positive selection pressure acting on four of the 26 coeliac disease loci: *IL18RAP* (rs7559479), *CCR1/5* (rs13098911), *SH2B3* (rs653178) and *PLEK* (rs17035378) (Table S4)[16-18]. We found the strongest selection signal (Fst=0.251) at the *SH2B3* locus, which was previously suggested to play a role in bacterial infection[16]. The *SH2B3* locus does not replicate in our north Indian cohort, possibly because of the selection pressure leading to a significant decrease in the risk allele frequency in north Indians (11% versus 47% in the Dutch).

Here we have replicated coeliac disease associations in a north Indian population and also localized and fine-mapped the association signals. However, signatures of positive selection pressure acting on 15% of the loci and our partial replication success may indicate that there are different mechanisms of disease pathogenesis in the two populations. The modest size of our cohort requires our results to be followed-up in more extensive samples. We recommend that in any future replication studies in north Indians or other non-Caucasian populations, both the European index SNPs, as well as the top variants from this study, should be tested to better distinguish relevant signals. Furthermore, our studies will benefit greatly from the 1000 Genomes sequence information for north Indians. This data will allow us to collate information on all the variants that are in LD with the transferable markers identified in our study and to describe more precisely the LD structure and genetic boundaries of variants strongly correlated with associated

SNPs. The key to effective fine-mapping is to perform dense genotyping that includes all known markers present in the north Indians. The results could have an immediate impact by allowing us to refine the associated coeliac disease regions to precise genetic intervals which include the causal variants.

MATERIALS AND METHODS

Ethical statement

Ethical approval for this study was granted by the respective institutional and university ethical committees. Informed written consent was acquired from all participants.

Study populations

We analyzed two distinct populations: north Indians and a Dutch population. Indian patients were recruited from a regional hospital in Punjab, northern India. Indian controls included blood donors recruited from the same region as the cases and who tested negative for coeliac disease serology. Dutch cases were recruited from University Medical Centre Utrecht, Leiden University Medical Centre and VU Medical Centre, Amsterdam. A small proportion of samples was recruited via the Dutch patients' society ('Glutenonderzoek'). Coeliac patients were diagnosed according to standard clinical, serological and histopathological criteria, including small intestinal biopsy. The majority of samples included in the current study have been described in detail elsewhere[12].

DNA extraction and genotyping

The great majority of DNA samples came from blood, while a small proportion of Dutch cases and controls were derived from saliva. Samples were hybridized on the Immunochip, a custom-made Infinium chip with 196,524 markers. Genotyping was carried out according to Illumina's protocol at the genotyping facility, University Medical Centre Groningen. NCBI assembly hg18 was used to map to the genome (Illumina manifest file Immuno_BeadChip_11419691_B.bpm).

Genotype data quality control

We required samples to have a 98% call rate based on the 172,242 high quality, manually clustered SNPs. Individuals showing high relatedness (PI_HAT > 0.2), or discordant sex were removed. Markers with significant deviation from Hardy-Weinberg equilibrium ($p > 10^{-3}$) or a per-SNP call rate < 99% were removed from the final dataset.

Population outliers were identified by multi-dimensional scaling (MDS implemented in PLINK[21]) on 11,192 SNPs that were common (MAF>0.05), LD pruned (a window of 1000 variants, sliding by 10 SNPs at a pair-wise SNP-SNP correlation r^2=0.05) and shared between Immunochip and HapMap3 samples. MDS analysis was performed jointly with HapMap GIH (Gujarati Indians in Houston, Texas), CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), YRI (Yoruba in Ibadan, Nigeria) and CHB (Han Chinese in Beijing, China) samples to identify major population outliers and also locally, separately for Dutch and Indians, to ensure the cases-controls were matching. All outlying samples were excluded from the analysis.

Immunochip contains 3016 variants submitted for the replication of the 'Reading and math skills' GWAS. As these SNPs are unlikely to be confounded by the immunity signal (49 SNPs mapping within HLA were excluded), we used this SNP set as a null reference for calculating genomic inflation.

Statistical analysis

Logistic regression was performed using PLINK v1.07[21]. Because we observed population substructure in the north Indian cohort, the north Indian association test was corrected for this by including the first three components of the MDS analysis as covariates in the logistic regression. For both cohorts, gender was included as a covariate in the logistic regression. 26 non-HLA loci, associated with coeliac disease at genome-wide significance level[10], were the focus of this study. We employed two

strategies to test for replication of European signals in Indians.

Exact replication tested for association signals (p≤0.05) among the 26 index SNPs from Dubois *et al.*[10] (Table S1). Two of the index SNPs were not present on Immunochip and were replaced with their best proxy; within the *IL18RAP* locus, rs917997 was replaced by rs7559479 (r^2=1) and within the *ITGA4/UBE2E3* locus, rs13010713 by rs1018326 (r^2=0.9).

Transferability of European signals in Indians was assessed by extracting all correlated markers, measured by r^2>0.05[22] between index SNP and all variants present within coeliac loci in CEU (from 1000 Genomes data). LD boundaries were defined by extending 0.1 cM to the left and right of the European focal SNP as defined by the HapMap3 recombination map (reported in Dubois *et al.*[10]). We then mapped the transferable variants to Immunochip and required at least one variant per locus at p<0.01 to replicate the European signal. The SNP tagging *ICOSLG* locus was not present in the 1000 Genomes data, so we used rs2838531 from HapMap3, a perfect proxy (r^2=1).

To assess the rate of transferability expected by chance, we permuted case-control labels and mapped the transferable SNPs into the permuted data. For 1000 permutations, 31 loci transferred at p<0.01.

Genotype clusters for all the significant markers identified by any of the three approaches were manually inspected in GenomeStudio.

GWAS and 1000 Genomes data

We used previously published coeliac disease GWAS[10] (Illumina Human Hap550 platform) and extracted 1,284 SNPs mapping in the 26 coeliac loci and shared with Immunochip. To assess the transferability rate we followed the same analysis flow as for the Immunochip data.

To estimate the coverage of the genetic variation within coeliac loci we used 1000 Genomes[23] data (May 2011, SNP calling) and extracted all annotated variants present in the sequenced samples of European descent: CEU (CEPH individuals), TSI (Tuscan individuals, Italy), FIN (HapMap Finnish individuals from Finland) and GBR (British individuals from England and Scotland). The May release of 1000 Genomes did not include any samples of Indian origin.

Test of heterogeneity

The homogeneity of ORs between Dutch and Indian cohorts was assessed by the Breslow-Day test. To estimate the accurate significance threshold for this test we permuted the disease status labels in both cohorts and calculated the 5% quantile of the nominal p-value distribution of Breslow-Day test statistics for 15,851 polymorphic variants in 26 coeliac loci among 1000 permutations.

Comparative LD background evaluation

For 11 transferable loci we visualized the LD structure for the entire regions in Indian and Dutch controls using Haploview v4.2[24]. No minor allele pruning was performed to get an accurate account of the cross-population LD differences. In all but three loci, we used all 736 north Indian controls and 1150 Dutch controls. Due to computational limitations for the *CCR1/5, IL2-IL21* and *ITGA4/UBE2E3* loci, we used 400 random controls from each population.

Estimation of fixation index (Fst)

A pair-wise fixation index[25] was calculated for all 12 European index SNPs. We used a control dataset from Indian, Netherlands, UK, Polish, Italian and Spanish populations. We calculated the significance level at Fst=0.079 based on the upper 5th percentile of Fst distribution for 8007 'neutral' SNPs. We determined neutral SNPs as intronic and uncorrelated with the coding variants (pair-wise r^2=0.05).

Software and resources

Raw genotype data was processed with GenomeStudio_v2010.3. Statistical analysis and quality control was performed with PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/

plink/). Plots were generated in R (http://www.r-project.org/). Haploview v4.2 was used to visualize LD plots. 1000 Genomes data was retrieved from http://www.1000genomes.org/. World-wide allele frequency distribution data was taken from (http://hgdp.uchicago.edu/).

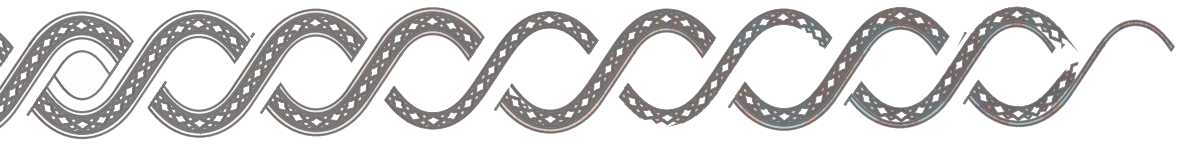
Competing interests
None to declare.

ACKNOWLEDGEMENTS
We thank the Dutch clinicians for recruiting celiac disease patients to give blood samples, as described in our previous studies (C.J. Mulder, G.J. Tack, W.H.M. Verbeek, R.H.J. Houwen, J.J. Schweizer). We thank the genotyping facility of the UMCG and Mathieu Platteel, Karin Fransen and Mitja Mitrovic for helping generate part of the Immunochip data and S. Jankipersadsing and A. Maatman at UMCG for preparing the samples.

Funding was provided by grants from the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative, and partially funded by the Dutch Government (BSIK03009 to C.W.) and the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.). We thank Jackie Senior for critically reading the manuscript.

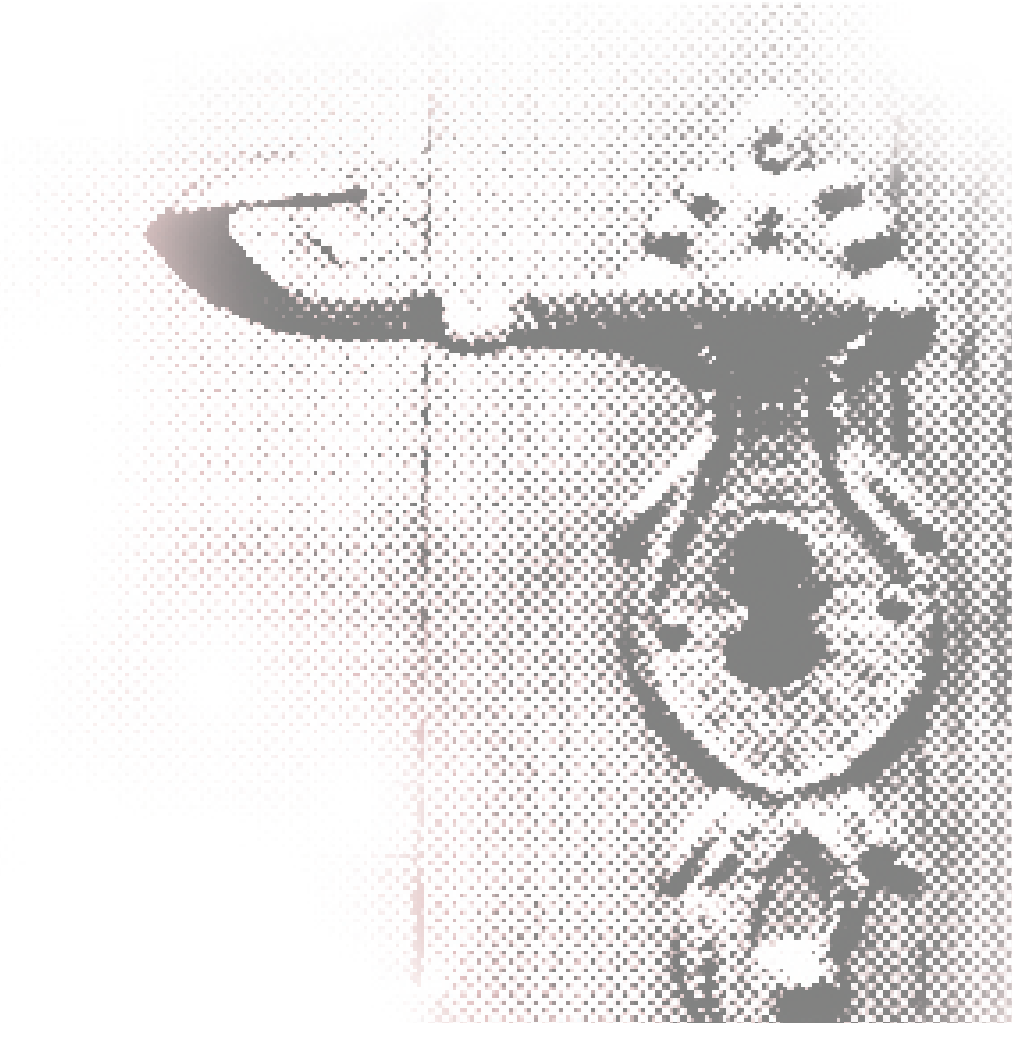
References

1. Manolio TA. (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363(2): 166-176.
2. Bustamante CD, Burchard EG, De la Vega FM. (2011) Genomics for the world. Nature 475(7355): 163-165.
3. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466(7307): 707-713.
4. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, et al. (2010) Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. PLoS Genet 6(8): e1001078.
5. Sim X, Ong RT, Suo C, Tay WT, Liu J, et al. (2011) Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from southeast asia. PLoS Genet 7(4): e1001363.
6. Fu J, Festen EA, Wijmenga C. (2011) Multi-ethnic studies in complex traits. Hum Mol Genet 20(R2): R206-13.
7. **Shriner D,** Adeyemo A, Gerry NP, Herbert A, Chen G, et al. (2009) Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. PLoS One 4(12): e8398.
8. Abadie V, Sollid LM, Barreiro LB, Jabri B. (2011) Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu Rev Immunol 29: 493-525.
9. Makharia GK, Verma AK, Amarchand R, Bhatnagar S, Das P, et al. (2011) Prevalence of celiac disease in the northern part of India: A community based study. J Gastroenterol Hepatol 26(5): 894-900.
10. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42(4): 295-302.
11. Cortes A, Brown MA. (2011) Promise and pitfalls of the immunochip. Arthritis Res Ther 13(1): 101.
12. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet (in press).
13. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, et al. (2011) Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am J Hum Genet 88(1): 57-69.
14. Need AC, Goldstein DB. (2009) Next generation disparities in human genomics: Concerns and remedies. Trends Genet 25(11): 489-494.
15. **Reich D**, Thangaraj K, Patterson N, Price AL, Singh L. (2009) Reconstructing Indian population history. Nature 461(7263): 489-494.
16. Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, et al. (2010) Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am J Hum Genet 86(6): 970-977.
17. Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, et al. (2002) A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A 99(16): 10539-10544.
18. Tang K, Thornton KR, Stoneking M. (2007) A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5(7): e171.
19. Corona E, Dudley JT, Butte AJ. (2010) Extreme evolutionary disparities seen in positive selection across seven complex diseases. PLoS One 5(8): e12236.
20. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, et al. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. N Engl J Med 359(26): 2767-2777.
21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3): 559-575.
22. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4(7): e1000130.
23. 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319): 1061-1073.
24. Barrett JC, Fry B, Maller J, Daly MJ. (2005) Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21(2): 263-265.
25. Weir BS CC. (1984) Estimating F-statistics for the analysis of population structure. Evol 38: 1358-1370.



Discussion and future perspectives

Chapter 8



Discussion

Part 1. From genetics to disease biology

In 2007 the first genome-wide association study (GWAS) for coeliac disease pointed to association with variants in a locus harbouring *IL2* and *IL21* genes[1]. This GWAS used ~300,000 single nucleotide polymorphisms (SNPs) genotyped in a small British cohort of 778 coeliac cases and 1,422 controls. Soon afterwards, this locus was replicated in rheumatoid arthritis and type 1 diabetes, pointing to a general risk locus for autoimmune diseases[2]. The top 1,020 most strongly associated variants from the coeliac disease GWAS were further taken for replication in an additional 1,643 cases and 3,406 controls and seven additional risk loci were identified[3]. As the majority of these loci harboured immune system-related genes, this pointed to an impaired immune response in coeliac disease involving both the adaptive as well as the innate immune system. At the same time, more associations for immune-mediated diseases were emerging and indicated overlap across the disease loci. This suggested that cross-disease replications could yield further insight into the genetics of diseases. For example, coeliac disease and type 1 diabetes are two diseases with a strong inflammatory component which co-segregate in populations and families, suggesting a common genetic background. A replication across these two disorders identified seven shared regions, suggesting that common biological mechanisms may lead to the aetiological features of immune-related diseases[4].

Three major conclusions have emerged from these studies: (1) SNPs scattered throughout the genome are a powerful gene-discovery tool when genotyped simultaneously in hundreds of case-control subjects, (2) the observed effects of associated variants are small and therefore large sample sizes (several thousands) are required to discover additional risk loci, and (3) immune-mediated diseases share a part of their genetic background. In this thesis, we aimed to further investigate the genetic background of coeliac disease by identifying additional genetic risk loci (Part 1) and dissecting the genetic architecture at established risk loci (Part 2). To identify more common disease-alleles we took three approaches: (i) deep replication of the UK coeliac disease GWAS in multi-population cohorts from Europe, (ii) meta-analysis of GWAS conducted across four European populations, and (iii) cross-disease replication. To interpret the coeliac disease associations we performed: (i) fine-mapping at these loci using a dense genotyping platform, and (ii) cross-ethnic mapping.

Through the genetic studies outlined in part 1 of this thesis we established associations between coeliac disease and 26 non-HLA loci. These are highly enriched for immune-system genes. For instance, the deep replication of the UK GWAS (Chapter 2) identified two loci, harbouring *REL* and *TNFAIP3* genes in proximity to associated SNPs, pointing towards the involvement of the nuclear factor kappa B (NF-κB) pathway. This was the first indication of the role of genetic variation in this immune pathway. Additional immune pathways emerged from the GWAS we conducted on 15,283 samples (Chapter 4), where we found association to loci harbouring the *THEMIS*, *ETS1* or *RUNX3* genes. These indicated the thymocyte selection in the thymus as an impaired pathway in coeliac disease. Some studies have suggested that high number of rotavirus infections may increase the risk of coeliac disease5. Our GWAS results point to the genetic basis for this increased susceptibility. The association of loci, including genes such as *TLR7/TLR8, BACH2* and *IRF4*, indicates an altered innate response to infection. Although the majority of coeliac disease SNPs map in the proximity of immune genes, there are loci which harbour genes with a more structural function, e.g. the *LPP* region. Very often we are biased when interpreting GWAS results and we tend to favour particular genes or pathways, hence neglecting genes with a less obvious function for the disease pathogenesis. Nonetheless, these "less obvious" coeliac disease genes could well implicate new venues of disease pathogenesis, e.g. the impaired barrier in the small intestine.

As more GWAS across diseases with immune or inflammatory components were conducted, it became clear that these diseases share a substantial part of their genetic background outside the well-known HLA region[6]. There is a large overlap between coeliac disease loci and those for type 1 diabetes, rheumatoid arthritis, multiple sclerosis or inflammatory bowel disease. In addition, there is also genetic sharing between coeliac disease and haematological, metabolic traits, and several types of cancers (Chapter 5). A proportion of 'disease-specific' genes could reflect distinctive disease biology, although this should be interpreted with caution. For example, the *LPP* region is the most associated non-HLA coeliac disease locus. Its association was established in 20083 and despite the large number of GWAS conducted in multiple immune- and other traits, it remained 'specific' for coeliac disease. Only two years later the *LPP* region was reported as a strong risk factor for vitiligo -immune disease not yet studied with GWAS[7]. This example shows that the genetic sharing will in fact be greater than currently observed, because of sample size limitations, platform coverage differences, and the stronger gene effect size in one disease than another. Therefore, cross-disease replication studies, where disease-alleles from one disorder are tested in the other, have emerged as an efficient approach to identify further risk genes at relatively low cost. In Chapter 3 we described such a study by cross-testing coeliac disease and rheumatoid arthritis variants. This pointed to six shared loci between the two diseases. Further genetic sharing can be identified via more comprehensive meta-analysis of GWAS results across diseases[8, 9]. The joint analysis of 15,283 samples from the coeliac GWAS and 22,770 samples from a rheumatoid arthritis study increased the genetic sharing from six to further fourteen regions[8]. In our Immunochip study (Chapter 6) we noted an excess of intermediate range p-values at the remaining 147 non-coeliac disease autoimmune loci, thus confirming the large genetic overlap between immune-mediated diseases and indicating hundreds of genes potentially underlying coeliac disease susceptibility.

It is nonetheless challenging to interpret the locus sharing. On the one hand, a consistent direction of the same allele in two diseases may imply a shared causal mechanism and the same risk genes. On the other hand, we often see the same loci but different SNPs, or the same SNPs but with opposite effects, are associated, e.g. the *IL18RAP* region shares the same SNP but with opposing effects for coeliac disease and type 1 diabetes[4]. It is therefore likely that a gene will protect from one disease while predisposing to another. Until extensive functional follow-up studies can be conducted to characterize the molecular mechanisms of associations and direct causal genes are identified, the interpretation of shared loci will remain unclear.

One way to gain insight into the molecular basis of genetic associations is to correlate the genotypes of disease-alleles with expression of the genes in their proximity (cis-eQTL mapping; further discussed in section 2.2 "Disease relevant tissues and target genes"). Such an approach may be especially helpful in the regions of strong linkage disequilibrium (LD) with multiple, functionally plausible disease genes. For example, SNP rs917997 maps to a region of strong LD at 2q12.1, covering four immune-related, strong candidate genes for coeliac disease, *IL18RAP, IL18R1, IL1RL1, IL1RL2*, nevertheless there is a strong cis-eQTL only for *IL18RAP*. Although this does not prove the causality it does give an indication of which could be the best gene for functional follow-up studies. Over 50% of coeliac disease SNPs are significant cis-eQTLs, indicating that subtle changes in gene expression are likely to be the underlying causal mechanism (Chapter 4). At successfully fine-mapped coeliac disease loci (Chapter 6), we observed clustering of associated variants to regulatory gene regions, either 3' or 5'. This further supports a deregulation of gene expression being the likely disease-driving mechanism (Figure 1).

Our Immunochip study[10], apart from fine-

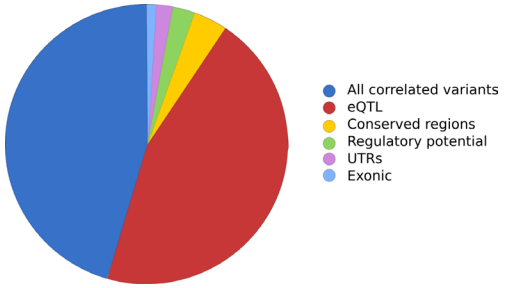


Figure 1 Some 4% of the variants map in the conserved regions, another 2% localizes in the regulatory regions in 3' or 5' UTRs, 3% map within the chromatin regions with high regulatory potential, and only 1% are exonic variants. Over half of coeliac disease-associated variants have an effect on gene expression (cis-eQTL).

mapping, also aimed at establishing new disease loci (Chapter 6). This platform includes ~200,000 SNPs with a strong pre-selection for immune system genes. It was designed to densely cover 183 loci associated with 12 immune-related diseases and included intermediate GWAS results for these diseases. This design made that Immunochip worked extremely well for coeliac disease, increasing the total number of risk loci from 26 to 39, covering 57 independent disease SNPs. However, despite establishing a large number of disease associated alleles, this study also indicated that a large proportion of heritability still needs to be explained. With 39 non-HLA loci we can only explain 14% of the genetic risk and further 40% is accounted for by the HLA locus. Even with 24,000 samples we had limited power to detect additional common, small effect size variants. The more disease loci we discover, the smaller the effects they account for. Hence, in order to detect these slight effects, we need even larger studies or further replications in extended case-control cohorts of maybe hundreds of thousands of individuals. Further common variations, possibly hundreds of them, are likely to cumulatively account for a large proportion of the disease risk.

Part 2. Future perspectives

Part 2.1. Identifying further genetic risk variants

The genetics of coeliac disease is now at an important crossroads. Tens of associations identified via GWAS are awaiting functional dissection in order to understand the mechanisms of causality, a step crucial for translating genetic findings into clinically relevant information. At the same time, hundreds of additional risk genes remain to be discovered. In this part, I discuss the future perspectives for coeliac disease in respect to further gene discovery. In most cases these approaches also apply to other complex traits, which are facing similar dilemma.

Heritability estimates

To carry on with gene mapping it is important to know how much of the genetic risk is explained by the established loci. We reported in Chapter 6 that the non-HLA variants account for 14% of the genetic variance, however, these results should be treated as estimates. The genetic variance was calculated based on assumptions of 1% disease prevalence and 50% heritability. To make accurate estimates of the genetic risk that can be attributed to variants identified via GWAS, it is essential to have accurate estimates of the heritability and disease prevalence.

First, coeliac disease has a broad spectrum of manifestation and therefore remains one of the most under-diagnosed diseases[11]. More accurate diagnosis will certainly influence prevalence estimates. Recent population-based studies in the UK indicate that the prevalence of coeliac disease is higher than previously assumed and it is now estimated at 1.6%[12]. Another study, based on a cohort of Swedish children, estimates coeliac disease prevalence at 3%[13]. Different disease prevalence impact the estimates of explained genetic variance, e.g. using a prevalence of 1.6% the associated non-HLA loci account for 16% of the heritability, whereas with 3% prevalence it is 19% (Figure 2).

Second, the heritability estimates may not be accurate. Italian twin studies calculated coeliac disease heritability at 89% assuming a population prevalence of 1/91 (1.1%).

However, this study used a small sample size, only 23 monozygotic and 50 dizygotic twin pairs, which resulted in a very broad range for the confidence interval, from 49%-100%[14]. Depending on the heritability assumed, 50% or 89%, the explained genetic variance accounted for by non-HLA alleles is 14% or 8%, respectively. This again varies for different disease prevalence (Figure 2), which all indicates the importance of accurate heritability estimates. To accurately calculate disease heritability, large-scale twin studies are required, especially to narrow down the confidence intervals. Alternatively, a method that estimates heritability based on genome-wide sharing between distantly related individuals could be used[15].

In summary, the explained variance should be treated as a rough estimate rather than a real measure. These calculations should be performed in a population-specific manner because: (i) the prevalence of the disease varies between populations, (ii) allele frequencies differ between populations, and (iii) heritability is expected to vary across environments. All these factors will affect the accuracy of estimates.

More common variants: large-scale GWAS and SNP prioritization for replication

A large number of identified coeliac disease-alleles only accounts for a small proportion of the disease risk, indicating that many more genes still remain to be discovered[16]. The GWAS we conducted, as well as most of the GWAS performed on complex diseases were underpowered to detect common variants of small effect sizes. The polygenic models of immune, complex diseases, Crohn's disease or type 1 diabetes[15], show that a substantial proportion of genetic variance is associated with common SNPs. Although coeliac disease was not tested in this model, it is highly likely

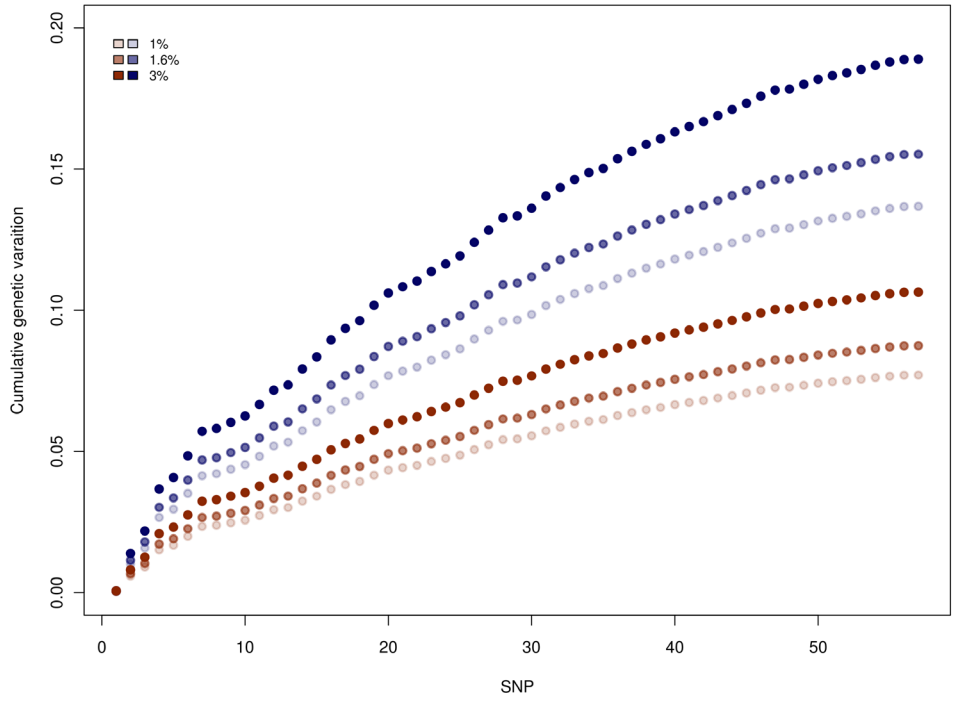


Figure 2 The genetic variance accounted for by the 57 non-HLA coeliac disease variants is strongly influenced by the heritability estimates (50% shown in blue and 89% in red) and by the different prevalence values. The extreme curves of the genetic variance differ by 11%, with 19% explained assuming 50% heritability and 3% disease prevalence but only 8% for 89% heritability and 1% prevalence.

that hundreds of genes will eventually account for its genetic background. GWAS employ strict significance thresholds when testing individual SNPs. This ensures high true-positive rates at the cost of increased false-negatives. In order to find additional risk variants at the genome-wide significance level, sample sizes of hundreds of thousands will be necessary. This is supported by observations from quantitative trait studies, such as for height and lipid levels[17], which both used over 100,000 individuals to provide the power to detect the small effects of common trait-alleles.

However, genome-wide genotyping of hundreds of thousands of individuals is very costly, and an alternative approach can be to prioritize SNPs for replication, e.g. based on their function or enrichment of the genes from similar pathways located in proximity to the associated SNPs. For example, the majority of coeliac disease loci harbour immune-related genes and thus the emerging pathways point towards different levels of immune-system regulation, e.g. the cytokine-cytokine receptor interaction pathway (Chapter 5). Replication of selected SNPs based on over-representation of immune pathways, by enrichment of SNPs in proximity to immune-related genes has proven successful for coeliac disease. First, there was the replication of >1,500 moderately associated GWAS SNPs that were enriched for immune-signal, which discovered ten new disease loci3 (Chapter 2). Second, more broadly by using the Immunochip10, which is strongly enriched for SNPs tagging immune system genes (Chapter 6), identified further thirteen loci.

Additionally, the analysis can be supplemented by pathway-based tools, which use prior biological knowledge on gene function to help more powerful analysis of the GWAS[18]. The principle for these approaches is to test if a group of related genes, in the same functional pathway, is associated with the disease. Pathway analysis approaches are enabling powerful association testing and helping formulate new hypotheses on disease aetiology. A successful example is rheumatoid arthritis, where SNPs were selected for replication based on their score in the Gene Relationships Across Implicated Loci (GRAIL)[19] method. This method performs statistical text mining of PubMed abstracts and scores the genes of interest for their functional relationships to the disease-established loci. Prioritization of SNPs for replication with this approach has established association with three additional loci[20].
Although pathway tools may aid in interpretation of GWAS results, or help generate hypothesis of disease pathogenesis and indicate additional gene candidates, they are not unbiased. When analysing a random selection of SNPs such tools especially favour the well-defined pathways[21]. Also, pathway tools may not consider those genes for which little is known about the function of their products. Association signals mapping to intergenic regions will also not be analysed. Although genes of poorly known function and intergenic regions are challenging to interpret and may not immediately fit into our understanding of disease pathogenesis, they should not be neglected as they may present new leads for the disease aetiology. And although pathway analysis can indicate the biological processes involved in the disease aetiology, the results should always be interpreted with caution; it is important that the results are followed-up by SNP replication to establish robust associations.

Another strategy may involve testing SNPs based on their function, e.g. a high percentage of coeliac disease alleles express an eQTL effect and one could then select for replication eQTL SNPs influencing expression of immune-related genes with moderate association in the GWAS results. Similar approach successfully identified an association between Crohn's disease and the *UBE2L3* gene[22].

The value of diverse ethnic groups
The prevalence of coeliac disease varies worldwide. From less than 0.2% in Germany[23] up to 5% in the Western Sahara[24]. In general, there is a lower prevalence of the disease in the developing countries. This could be partly

explained by different eating habits and a more diverse gluten dietary intake, which to some extent correlates with the disease rate. This is also supported by the Swedish coeliac disease epidemic. Between 1984 and 1996 Sweden experienced a marked epidemic of coeliac disease[13], reaching a disease incidence rate of 3%. It is thought this high rate can be partly explained by a change made in the infant feeding guidelines. Before the coeliac disease epidemic, it was recommended to introduce gluten in the infant diet from 4-6 months of age, at the same time breast-feeding was being discontinued. Independently, this coincided with the commercial baby cereal milk products having higher gluten content. The end of the epidemic was preceded by a recommendation to introduce gluten gradually while the baby was still breast-feeding and this also overlapped with a reduction in the gluten content in commercial infant food.

Yet, it is not clear whether coeliac disease is rare in some populations because of a lack of predisposing genes or whether the genes are not expressed because food containing gluten is not traditionally consumed. Studies on second-generation immigrants and adoptees in Sweden show a decreased incidence of coeliac disease among children whose biological parents are from countries with a low disease prevalence and at the same time children of parents from Western, Eastern and Northern Europe have a similar incidence of coeliac disease as in Sweden[25]. This suggests that ethnic genetic heterogeneity contributes to the worldwide variation.

Different ethnic backgrounds may emphasize different mechanisms of disease pathogenesis. Different environments may lead to different selection pressure, which impacts the allele frequencies. Four of coeliac disease loci (*SH2B3, IL18RAP, PLEK* and *CCR1/5*) have signatures of positive selection. At three loci the selection is acting on the protective allele. Interestingly, at the *SH2B3* locus the favoured risk-allele is not present at all in South Asians or South Africans but in Europeans it reaches almost 50% frequency. This suggests that new, important associations between disease and genetic variants may be found easily in populations with locally common risk allele frequencies. GWAS has proved very successful in relatively small and homogeneous populations, such as in Iceland or Sardinia. Limited migration and large families are common for populations in Africa or Latin America, and some groups in these continents exhibit high coeliac disease rates. New genetic insights will emerge from studying such diverse ethnicities, for example groups from Mexico, Uruguay, Algeria and Syria (see Figure 2 of Preface). In all these populations the prevalence of coeliac disease is at increased levels (>1.6%). Of particular interest is the population of the Western Sahara, where the coeliac disease rate reaches an especially high level of 5%.

The study of blood lipid levels shows that most of the common disease-variants contribute to the risk in other ethnic groups[17]. Yet this observation cannot be generalized as only about 4% of the GWAS were conducted on non-European populations2[6], and it is likely that findings from one population may not always translate to others. For example, rs9282541 is strongly associated with type 2 diabetes, obesity and low high-density lipoprotein cholesterol (HDL-C), but is exclusive to Native Americans and not present in European, Asian or African individuals[27].

When attempting to replicate associations established on Europeans in other ethnicities, it is important to account for cross-population differences in LD structure. GWAS identified variants are, with few exceptions, associated to, but not causing the disease. Depending on the LD block structure and frequencies of tag and causal genetic variants, associations from one population may not be generalized across another. Especially testing the index European SNP (reported SNP from a discovery study) may not be sufficient to replicate the signal in another ethnic group. It is then better to use the index SNPs jointly with variants in LD with it in replication studies[28]. Using this approach

we were able to replicate 13 of the 26 coeliac disease loci in a north Indian population. Had we tested the index SNP only, we would have missed signals for eight loci (Chapter 7). The choice of a north Indian population was beneficial because this population is ethnically different from Europeans, resulting in a lesser degree of long-range LD. This has proven advantageous for fine-mapping association signals (section 2.2, "Fine-mapping").

Finally, part of the "hidden heritability" is also attributed to low-frequency and rare variants, i.e. those present at MAF less than 5%. However, rare variants are population-specific[29], even more opting for studies which include diverse genetic populations, or else our understanding of the disease genetics will be strongly biased.

Rare variants

The genetic architecture of complex diseases comprises both common and rare variants. Part of the ongoing debate is whether the "hidden heritability" can to some extent be explained by the rare, large effect size variants. GWAS have been successful at finding associations to common variants, partly because GWAS platforms mainly consist of variants with MAF greater than 5%. Due to this and the fact that LD correlation between common and rare variants is limited, direct sequencing is required to uncover rare disease variants. Studies to identify rare variants in the proximity of the common variant signal have already shown success in e.g. type 1 diabetes[30].

In our Immunochip study[10] (Chapter 6) we did identified low frequency and rare variants (minor allele frequency MAF <5%) associated with coeliac disease but of moderate effect sizes; we did not observe any rare variant with very strong effect. The strongest effect had an odds ratio (OR) of 1.7 for a variant (imm_16_11281298) with a MAF of 0.004%. We also reasoned that if rare, large effect size variants underlie associations at coeliac disease loci, we would expect the MAF of most associated variants to decrease with increasing genotyping density. This was not the case, and compared to our GWAS we did not observe any such trend. However, our study had certain limitations.

Rare variants tend to be population-specific and require large sample sizes to reach a high significance level. Our Immunochip study is overrepresented by the UK population (16,000 UK samples in a total of 24,000). Variants with less than 5% frequency should be analysed in a population-specific manner as in the pooled analysis population specific effects may be diluted by other cohorts. However, except for the UK cohort, the other six cohorts were not sufficiently powered for this type of analysis. Furthermore, rare variant detection requires case resequencing data, as has been included on the Immunochip, but the sequencing was performed only on UK cases and only for thee coeliac loci. Therefore, again, if rare, population-specific variants confer the risk at coeliac disease loci they should be analysed only in the UK samples. In addition, the Immunochip has a patchy coverage of rare variants. It was designed based on an early release of the 1000 Genomes Project, which should identify the vast majority of common (MAF>5%) variants, but it was underpowered to comprehensively detect rare frequency variants. Our Immunochip dataset captured 47,602 variants with a MAF<5%. On the other hand, despite the uneven coverage of rare variants, we might have expected to capture at least one rare variant from the 39 loci if single rare variants explain the highest association signals, but we did not. It seems unlikely that a single rare variant explains most of the association signal. However, the statistical analysis we applied cannot robustly refuse that a combination of multiple rare variants jointly explains the association signal.

Part 2.2. From SNP association to function

Genome-wide association studies (GWAS), in conjunction with the most recent Immunochip study, have led to the identification of 39 genomic regions being associated with the risk of coeliac disease[1,3,8,10,31,32]. These associations

have improved our understanding of the disease biology and pointed to new disease pathways. Despite the exciting progress made with mapping disease loci, it is disappointing that for the majority of loci it is still unclear how the risk alleles alter the function of neighbouring genes to disrupt critical molecular pathways leading to disease. Without knowing more about the biological functions of the risk alleles, it is very difficult to develop novel therapeutic strategies to treat or possibly cure autoimmune diseases.

There are at least three reasons why it has been difficult to understand the biological function of risk alleles. First, risk loci often contain multiple genes, many of which may be biologically strong candidate genes, e.g. the 3p21.31 region, harbouring a cluster of chemokine receptors. And even more challenging, some risk loci map to intergenic regions where no genes are present. Second, multiple alleles may be in linkage disequilibrium (LD) with the disease-associated SNP. Unless a variant disrupts an obvious conserved sequence motif (e.g. by altering the protein-coding structure of a gene), it is not possible to determine which is the "causal variant". And third, without knowing the causal gene or causal variant, it is difficult to collate information across risk loci to gain insight into critical biological pathways.

Fine-mapping

Before understanding a function of the disease variant it is necessary to identify the most plausible causal variant(s) that can be further taken into functional follow-up studies. In majority GWAS establish associations between the disease trait and tag-SNPs. This is due to the design of the genotyping platforms used in GWAS, which try to capture most of the LD structure. Fine-mapping aims at narrowing the association signal down to the disease-causing variant, which is in LD with the associated, tag-SNP. It requires sequencing of cases and controls and can be supplemented by the 1000 Genomes project, which catalogues common (>5%) and low (1-5%) frequency variants in individuals from diverse ethnic populations[33]. Currently, intensive whole-genome sequencing efforts in hundreds of samples are being conducted and aim at generating comprehensive and detailed catalogues of the genetic variants specific to the populations of interest, for example the "Genome of the Netherlands" project. The "Genome of the Netherlands" will characterize the genomes of 750 Dutch individuals (250 trios) from four major provinces in the Netherlands. Such a reference set will have a higher specificity to complement analyses focused on the samples of Dutch origin than the 1000 Genomes CEU panel.

One can choose to follow-up only a subset of "best" plausible causal variants or to genotype all the identified sequence variants. The latter is a more unbiased approach and omits the chance of missing the causal variant in the replication, due to prioritization of the follow-up SNPs. A similar reasoning underlies the design of Immunochip, where all the variants, irrespective of the LD correlation or minor allele frequencies, were submitted for the design. This strategy provided a detailed, unbiased picture of the genetic architecture at the disease loci. Ideally one would like to perform a sequencing case-control association study to fully exclude the possibility of missing the causal variant.

Another fine-mapping approach involves using GWAS information from ethnically different populations, on the assumption that similar associations are generalized across these populations. One of the most genetically informative ethnic groups to study is the African population, which is much older than Europeans and therefore has a generally smaller LD block structure. This gives opportunities to localize association signals to a smaller genetic interval and narrow down the list of putative disease variants.

Evaluating functions of associated variants

HLA has long been studied in the context of coeliac disease pathogenesis and the molecular mechanism is well known. However, this is not

the case for the other loci, with the exception of the 12q24.12 region, where more than ten genes are present in the LD block but the strongest association maps to a coding SNP within the *SH2B3* gene. Individuals carrying the risk allele show an increased pro-inflammatory immune response against bacterial infection[34]. This locus is also under positive selection pressure. It is a very rare example for complex diseases, where the most significant GWAS SNP is a coding variant. In the other loci, only a few carry the coding variant in LD with the most associated SNP.

On the other hand we observe that most of the signals localize in the plausible regulatory regions in proximity to or within the 5′ and 3′ ends of transcription. This could indicate a disease mechanism via deregulation of gene expression. Consistent with that is the fact that for over 50% of coeliac disease variants in our GWAS study we observed correlation with gene expression (cis-eQTL). More broadly, genome-wide studies of gene expression linked with GWAS data have shown that approximately 40% of disease-associated SNPs affect expression levels of genes in cis[35].

The emerging picture thus suggests that many of the causal variants will disrupt the gene expression of a neighbouring causal gene, either by (i) affecting pathways that control activity of transcription factors[36], (ii) directly disrupting sequence motifs that bind transcription factors regulating the expression of neighbouring genes, or (iii) disrupting or creating miRNA binding motifs[37]. For example, NF-κB, one of the most important transcription factors modulating immune response, has been implicated in the pathogenesis of coeliac disease and multiple other immune-mediated diseases. Associations with genes that regulate the activity of NF-κB have been described (e.g. TNFAIP3 in psoriasis[38,39], rheumatoid arthritis[40, 41], coeliac disease[32] and systemic lupus erythematosus[42, 43], as well as associations mapped to a subunit of NF-κB itself, c-REL (ulcerative colitis[44], Crohn's disease[45], psoriasis[39], rheumatoid arthritis[41], and coeliac disease[32]).

It is relatively easy to assess the causality for the coding variants as they may alter protein sequences, but more challenging for the non-coding ones. Recently, a new method has become available to identify functional motifs in DNA sequence: chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) to generate genome-wide chromatin data sets. By mediating regulatory signals and controlling DNA accessibility, active chromatin elements steer gene expression in a tissue-specific manner. The ENCODE Project[46] provides comprehensive genome annotation of the regulatory elements that emerge from chromatin profiling studies. Chromatin regulatory elements include enhancers, insulators, promoters, and silencers, and of these, enhancers are the most abundant and often cell-specific[47]. Associated, disease variants can therefore be annotated if they map within such chromatin functional elements. Additionally, associated variants can also be tested for their consequences on regulatory motifs (disruption or creation) for predicted transcription factors. Given the high rate of eQTL effects among disease-associated SNPs, the enrichment of these variants within regulatory elements is expected. This can pinpoint the specific disease mechanisms by which a disease-allele leads to the phenotype.

Characterizing the landscape of disease-associated SNPs by integrating fine-mapping information with chromatin modification (e.g. chromatin regulatory elements, open chromatin), genome conservation and e-QTL data will provide an important step towards understanding how risk alleles affect function. With the emerging picture of noncoding variants contributing to the disease, it would be desirable to design a chip that combines sequencing variants with chromatin regulatory information to aid fine-mapping efforts.

Disease-relevant tissues and target genes

While many individual loci have emerged from GWAS, identification of the

exact causal gene and tissues in which they act has remained a challenge. In particular, autoimmune diseases are a major challenge, with a broad range of immune-cell types that may be impacted by genetic variation and lead to the disease. Thus it is essential that the function of genetic variants is studied in systems most representative for the disease. It is not straightforward to pinpoint the right disease tissues, however few approaches have recently been suggested, e.g. by scoring the enrichment of genes implicated by disease-associated SNPs across different cell types. For instance, there is a significant enrichment of the gene expression of CD_4+ effector memory T-cell genes among rheumatoid arthritis loci[48].

Another approach could involve annotating the disease-associated variants. For example, chromatin enhancers have high tissue specificity and mapping associated variants for chromatin functional elements shows that disease SNPs are enriched within these enhancer elements[49]. Further, testing the enrichment of enhancers with disease-alleles across different cell types may explain the tissue- and disease-specific nature of the disease-variants. For example, systemic lupus erythematosus SNPs are enriched within strong enhancer states in B-lymphoblastoid cells, but not in eight other cell types tested[49]. This, however, requires a comprehensive catalogue for many different cell types. For coeliac disease-alleles, it would be important to test not only immune cells but also cells making up the small intestine.

After choosing the disease-relevant tissue, the next step is to identify the disease-causing gene. A considerable proportion of associated loci will harbour variants that influence the abundance of specific transcripts. Studying the correlation between genetic variants and the expression of neighbouring genes is a straightforward way to link risk-alleles with putative target genes. Importantly, not only messenger RNA (mRNA) but also microRNA (miRNA) and non-coding (ncRNA) transcripts can be regulated by genetic variation and should be taken into account in such an analysis. A complementary approach to define the local impact of genetic variation on gene expression is allele-specific gene expression in individuals heterozygous for risk alleles. Nonetheless, an identified transcript that strongly correlates with a disease-variant does not mean it is definitely involved in the disease and functional follow-up studies will be critical for proving the causality.

Environmental risk factors and genetic variation
Our understanding of the complex interplay between genetics and environment poses a particular challenge for post-GWAS follow-up. Determining the disease-triggering factors is crucial for future studies and only when we identify the environmental factors will we be able to link them with genetic information and gain a full picture of the disease pathogenesis. This picture will help us to understand the molecular mechanisms of the genetic associations. Coeliac disease is an excellent example of such a translation of information, the strongest genetic determinant is HLA, more specifically the *HLA-DQ2* and *HLA-DQ8* molecules, which bind the gluten peptides, the environmental factor. This knowledge has immediate translation to the patients, as restricting the presence of gluten in the diet completely reverses the disease symptoms for 95% of coeliac patients. To gather detailed environmental information, prospective studies including large numbers of participants are required. Such efforts are ongoing, for example in the Netherlands there is "LifeLines", a three-generation population-based project[50]. LifeLines is recruiting 165,000 participants from the northern provinces of the Netherlands, collecting extensive phenotype information and generating genotyping profiles to understand how the interplay between genetic and non-genetic factors modifies an individual's susceptibility to multifactorial diseases.

Of particular interest is how diet influences the composition and dynamics of the gut microbial communities (microbiota) and impacts the innate and adaptive immune system. The digestive tract is the major part of the immune system and highly exposed to environmental

factors. Our knowledge of the species and composition of the human gut microbiome has been greatly expanded in recent years. It now appears that, across the world, human populations share three robust clusters, composed of multiple microbial species, and there are a limited number of symbiotic, well-balanced host-microbial states, which may have different responses to diet or drugs[51]. The microbial composition strongly correlates with human phenotypes, for instance there is a significant difference between the microbiomes of obese and lean people, indicating the diagnostic potential of microbial markers[52]. It could be of great diagnostic value to analyse, and if possible, create the microbial profile for coeliac disease. Further, the integration of microbiome profiles with genetic information could become an important pre-diagnostic tool for the diseases.

Part 2.3 Clinical translation

Despite many genes being identified for complex traits their use for clinical testing is still poor. First, the majority of identified variants are surrogates of the real disease-alleles and their effect-sizes are only proxies of the real effects. Second, the identified variants account for very modest effect sizes. Coeliac disease is one of the genetically best understood complex diseases, with its very strong risk effect of the HLA molecules. Due to this, prediction models for coeliac disease are some of the most successful to identify individuals at high disease risk. If individuals carry an intermediate HLA risk, the load of non-HLA risk alleles improves the prediction model[53]. This model will improve further, as we identify more common and rare variants, and will help to better calculate the risk for individual to develop the disease. With the genetic advances in coeliac and other immune-related disease, it will be possible to generate a general profile of the risk for immune-related diseases.

Although current GWAS results still lack high predictive value, identified loci can point towards novel pathways and therefore lead to identification of therapeutic targets. Genetic overlap between immune-mediated diseases indicates shared molecular pathways, which may provide new targets for clinical treatment, while similar compounds could be used for a broad spectrum of diseases with respect to an individual's genetic profile. GWAS results can also point towards environmental factors contributing to the disease, enabling public-health prevention measures to be taken.

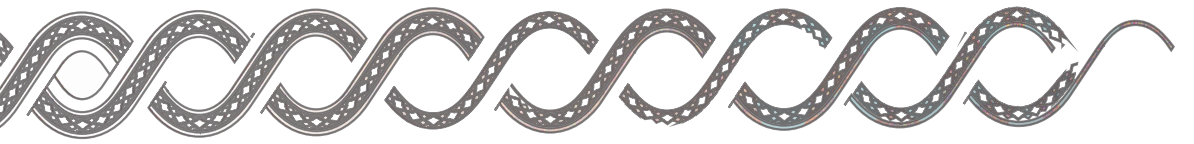
GWAS results have the potential to facilitate personalized medicine and clinical care but there is still a long way to go. Until we identify the disease risk alleles, understand their molecular consequences on a single locus level as well as on the biological system level, we will not be able to effectively translate the genetics to clinical care. Ultimately, the combination of genetic markers and biomarkers of the disease, for example, antibodies or microbiome signatures, may lead to better pre-diagnosis and effective prevention.

REFERENCES

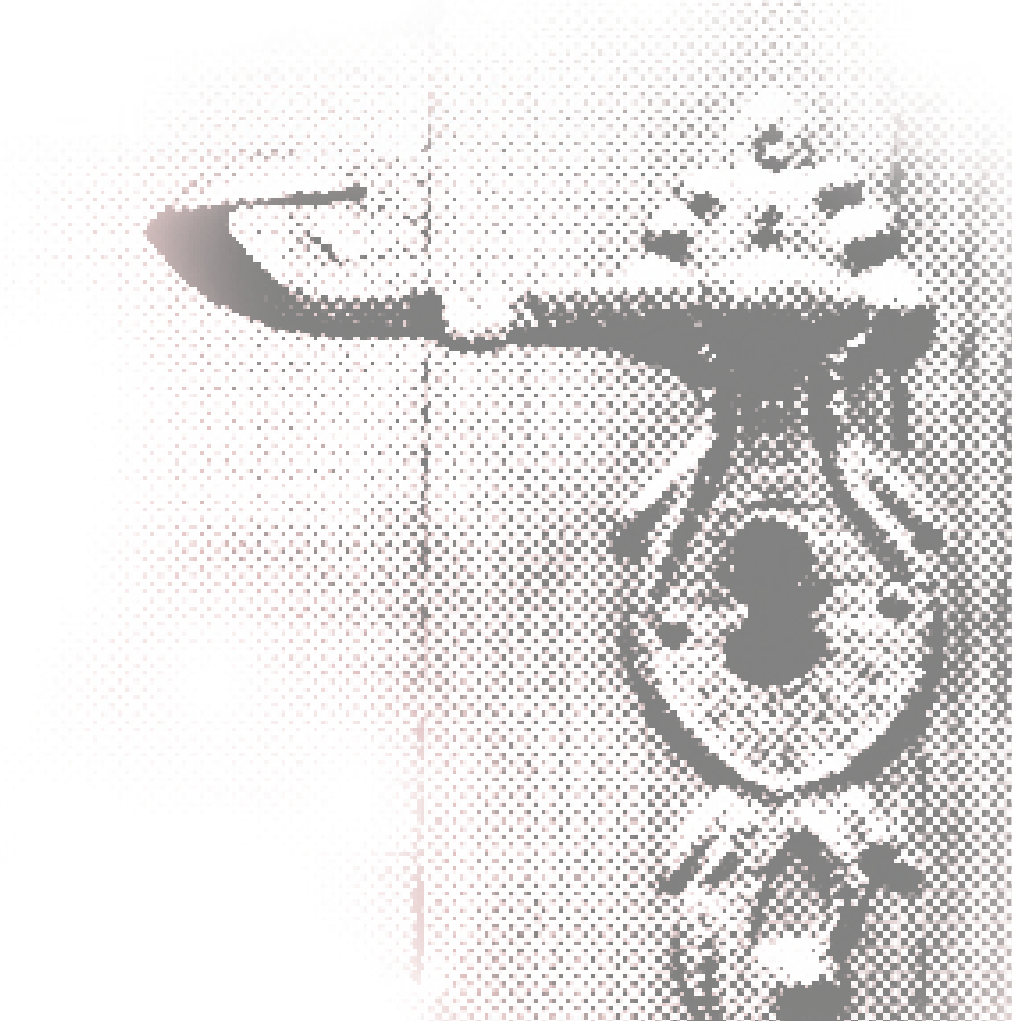
1. **van Heel, D. A**. et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827-829 (2007).
2. **Zhernakova, A**. et al. Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. Am. J. Hum. Genet. 81, 1284-1288 (2007).
3. **Hunt, K. A**. et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40, 395-402 (2008).
4. **Smyth, D. J**. et al. Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med. 359, 2767-2777 (2008).
5. **Stene, L. C**. et al. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am. J. Gastroenterol. 101, 2333-2340 (2006).
6. **Zhernakova, A.**, van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet. 10, 43-55 (2009).
7. **Jin, Y.** et al. Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. N. Engl. J. Med. 362, 1686-1697 (2010).
8. **Zhernakova, A.** et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet. 7, e1002004 (2011).
9. **Festen, E. A**. et al. A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. PLoS Genet. 7, e1001283 (2011).

10. **Trynka, G**. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. (2011).
11. **Green, P. H**. & Cellier, C. Celiac disease. N. Engl. J. Med. 357, 1731-1743 (2007).
12. **Walker, M. M**. et al. Detection of celiac disease and lymphocytic enteropathy by parallel serology and histopathology in a population-based study. Gastroenterology 139, 112-119 (2010).
13. **Myleus, A.** et al. Celiac disease revealed in 3% of Swedish 12-year-olds born during an epidemic. J. Pediatr. Gastroenterol. Nutr. 49, 170-176 (2009).
14. **Nistico, L.** et al. Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55, 803-808 (2006).
15. **Lee, S. H.**, Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. 88, 294-305 (2011).
16. **Manolio, T. A**. et al. Finding the missing heritability of complex diseases. Nature 461, 747-753 (2009).
17. **Teslovich, T. M**. et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-713 (2010).
18. **Wang, K**., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. 11, 843-854 (2010).
19. **Raychaudhuri, S.** et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 5, e1000534 (2009).
20. **Raychaudhuri, S**. et al. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet. 41, 1313-1318 (2009).
21. **Elbers, C. C**. et al. Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet. Epidemiol. 33, 419-431 (2009).
22. **Fransen, K.** et al. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Hum. Mol. Genet. 19, 3482-3488 (2010).
23. **Abadie, V.**, Sollid, L. M., Barreiro, L. B. & Jabri, B. Integration of genetic and immunological insights into a model of celiac disease pathogenesis. Annu. Rev. Immunol. 29, 493-525 (2011).
24. **Catassi, C.** et al. Why is coeliac disease endemic in the people of the Sahara? Lancet 354, 647-648 (1999).
25. **Ji, J.**, Ludvigsson, J. F., Sundquist, K., Sundquist, J. & Hemminki, K. Incidence of celiac disease among second-generation immigrants and adoptees from abroad in Sweden: evidence for ethnic differences in susceptibility. Scand. J. Gastroenterol. 46, 844-848 (2011).
26. **Need, A. C. & Goldstein, D. B**. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 25, 489-494 (2009).
27. **Acuna-Alonzo, V.** et al. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Hum. Mol. Genet. 19, 2877-2885 (2010).
28. **Shriner, D.** et al. Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. PLoS One 4, e8398 (2009).
29. **Gravel, S.** et al. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A. 108, 11983-11988 (2011).
30. **Nejentsev, S.,** Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324, 387-389 (2009).
31. **Trynka, G.** et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. Gut 58, 1078-1083 (2009).
32. **Dubois, P. C.** et al. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295-302 (2010).
33. **1000 Genomes Project Consortium.** A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073 (2010).
34. **Zhernakova, A.** et al. Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am. J. Hum. Genet. 86, 970-977 (2010).
35. **Fehrmann, R. S.** et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7, e1002197 (2011).
36. **Adrianto, I.** et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat. Genet. 43, 253-258 (2011).
37. **Brest, P.** et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat. Genet. 43, 242-245 (2011).
38. **Nair, R. P.** et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet. 41, 199-204 (2009).
39. **Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2** et al. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat. Genet. 42, 985-990 (2010).
40. **Plenge, R. M**. et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat. Genet. 39, 1477-1482 (2007).
41. **Stahl, E. A.** et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42, 508-514 (2010).
42. **Graham, R. R.** et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat. Genet. 40, 1059-1061 (2008).
43. **Han, J. W**. et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet. 41, 1234-1237 (2009).
44. **McGovern, D. P**. et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat. Genet. 42, 332-337 (2010).
45. **Franke, A.** et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42, 1118-1125 (2010).
46. **ENCODE Project Consortium** et al. Identification and analysis of functional elements in 1% of the human

genome by the ENCODE pilot project. Nature 447, 799-816 (2007).
47. **Visel, A**. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457, 854-858 (2009).
48. **Hu, X.** et al. Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. Am. J. Hum. Genet. (2011).
49. **Ernst, J.** et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-49 (2011).
50. **Stolk, R. P.** et al. Universal risk factors for multifactorial diseases: LifeLines: a three-generation population-based study. Eur. J. Epidemiol. 23, 67-74 (2008).
51. **Arumugam, M.** et al. Enterotypes of the human gut microbiome. Nature 473, 174-180 (2011).
52. **Turnbaugh, P. J.** et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027-1031 (2006).
53. **Romanos, J.** et al. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology 137, 834-40, 840.e1-3 (2009).



Summary



Genetics is one of the most dynamically evolving fields of science. The full sequencing of the human genome was completed only in 2003. It was an immense effort which took 13 years to perform, at a cost of $3,000,000,000. Currently the cost of sequencing a single genome is $10,000 and it takes less than a month. From the sequence of the human genome we learned that there are about 10,000,000 common, genetic variations, changes at a single position in the DNA, these are called single nucleotide polymorphisms (SNPs). The majority of SNPs are expected to be neutral, however, some of them are present with a higher frequency in patients suffering from a certain disease when compared to healthy individuals. Recent technological advances have led to platforms capable of simultaneously testing (genotyping) hundreds of thousands of SNPs (from 300,000 up to 2,000,000) in hundreds of individuals. These platforms are used in case-control studies and allow us to infer the genotypes of each SNP and compare the frequencies between patients and healthy individuals. Statistically significant differences indicate the regions of the genome that are relevant for the disease development, disease loci. Such a study design is called a genome-wide association study (GWAS). However, one of its drawbacks is that it does not identify the causal gene or mutation directly, but rather points to a region harbouring the disease gene. Often such regions are large and they contain multiple genes which are strongly correlated with each other due to the linkage disequilibrium (LD) structure. It therefore becomes difficult to pinpoint the true disease-causing gene and after identifying the disease regions with GWAS, we need to perform further fine-mapping studies to narrow down the signal and uncover the casual variant.

It is due to this technological development that the past few years have yielded such a tremendous advance in our understanding of the genetic background of many diseases. The success has been most pronounced for monogenic disorders, in which a change in a single gene leads to the development of the disease. About five years ago we also began to see successes for complex diseases. These are disorders that result from small, simultaneous dysfunctions of many, even hundreds, of genes. Yet, it is not simply the genetic component that leads to development of the disease, but also the interplay between disease genes and one or more environmental factors.

Coeliac disease is the most common food intolerance, affecting some 1-3% of Western populations. It results from an intolerance to gluten peptides in genetically susceptible individuals. Gluten is found in wheat, barley and rye, and is widely present in Western food products. The intake of gluten causes a strong immune response and leads to damage of the small intestine in such individuals. People with coeliac disease can suffer from a broad spectrum of symptoms, from classical ones such as diarrhoea, abdominal distension and abdominal pain, to atypical ones that include anaemia, osteoporosis and neurological symptoms.

Genetically, coeliac disease is well characterized by the strong effect of two particular HLA molecules, HLA-DQ2 and HLA-DQ8. These molecules account for 35-40% of the genetic risk, yet are present in 30% of the general population, indicating that other genetic risk factors must be involved. By 2008, eight non-HLA regions had been identified by a GWAS in coeliac disease patients. The aim of this thesis was to further identify the genes predisposing to the disease development.

In Chapter 2, we describe the results of deep replication studies of the coeliac disease GWAS. We genotyped approximately 500 SNPs that had a moderate association signal in the GWAS in 1,682 cases and 3,258 controls and identified association of two additional genomic regions. Both regions harboured genes involved in modulating the activity of the NF-κB pathway, one of the essential transcription factors mediating the immune response.

Immune-related diseases often co-occur in families or patients, for example, patients with

coeliac disease are also often affected by type 1 diabetes. In Chapter 3 we analyse the extent of genetic sharing between two autoimmune conditions, coeliac disease and rheumatoid arthritis. We tested if the genes associated to one disease also confer risk for the other. This strategy enabled us to report a total of six shared regions, confirming two previously known loci and identifying four new ones.

Further coeliac disease genes were discovered via a large-scale GWAS (described in Chapter 4) in 15,000 individuals from four populations (UK, the Netherlands, Italy and Finland). The top 131 SNPs that showed the strongest association in the analysis across these populations were replicated in an additional 10,000 individuals from USA, Hungary, Ireland, Poland, Spain, Italy and Finland. With this approach, we brought the total number of coeliac disease loci to 26. The majority of these regions harbour genes with an immune-related function. Apart from the genes involved in modulating the innate and adaptive immune responses, we also describe genes implicating the role of lymphocyte development on the thymus, as well as genes involved in the innate immune detection of viral infection. Importantly, we report that over 50% of coeliac disease-associated SNPs have a strong effect on gene expression. This suggests that one of the main causal mechanisms for coeliac disease associations, and probably for other complex diseases, is the alteration caused in the levels of gene expression.

In Chapter 5, we give an overview of the progress made by genetic studies in coeliac disease and of the advance in our understanding of the molecular mechanisms underlying the disease.

Our Immunochip study (described in Chapter 6) further confirmed that one of the major mechanisms driving the disease pathogenesis is the deregulation of gene expression. The Immunochip is a custom-made platform designed specifically for use in genetic studies of immune-related diseases. Its SNP content has been enriched for regions harbouring 'immune genes' and SNPs previously associated to any of the immune-related diseases. In particular, the Immunochip was designed to aid fine-mapping, to more precisely localize the association signal, in 183 regions associated to immune-related disorders. For this work, these regions have been enriched with markers at a much greater density than in GWAS platforms (10-20x). The greater SNP density provides a better resolution of the linkage disequilibrium and permits a more precise localization of the association signal. The application of the Immunochip in 12,000 coeliac disease patients and 12,000 controls from six different countries successfully allowed the signal to be localized to a single gene at more than half of the coeliac disease loci. Furthermore, at 20% of the coeliac loci, the signal was localized in a gene regulatory region, again pointing towards the deregulation of gene expression as a molecular mechanism underlying the association signals. In addition to fine-mapping, the Immunochip study identified 13 additional coeliac disease loci, making a total of 39 non-HLA regions carrying a coeliac disease risk.

To further understand the genetic architecture of coeliac disease, we performed a cross-ethnic analysis, between the Dutch and northern Indians. We successfully replicated 13 of 26 tested loci. We observed a difference in long-range linkage disequilibrium between these populations, which resulted in finer signal localization in the northern Indians at five loci and a signal shift at the remaining ones. For a number of loci, we were able to report signatures of positive selection pressure acting upon them.

The main conclusions from this thesis are:

- 39 non-HLA loci have been associated with coeliac disease so far and, together with HLA, they explain 53% of the genetic risk.
- Coeliac disease and other immune-related diseases share substantial parts of their genetic background.
- Loci associated to coeliac disease are greatly enriched for immune-system genes.
- Over half of the associated variants affect the

expression of neighbouring genes.
- Coeliac disease-associated variants are mostly non-coding and localize in the gene regulatory regions, indicating that gene deregulation can be one of the major molecular mechanism driving the disease pathogenesis.
- 13 coeliac disease loci mapped in European populations also confer coeliac disease risk in a north Indian population.

Glossary

DNA: the molecule that stores the genetic information. The basic unit building up the DNA is a nucleotide.

Genome: the hereditary information describing the development and functioning of living organisms. It consists of genes and non-coding parts of the genome.

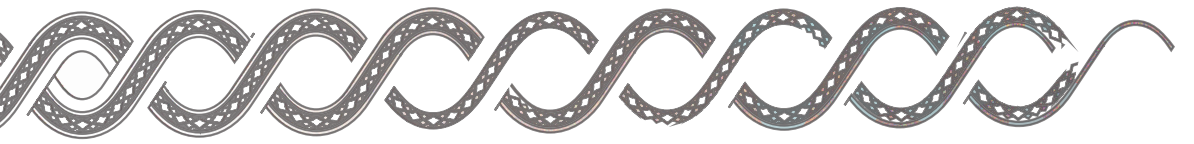
GWAS: genome-wide association study, the genotyping of hundreds of thousands of SNPs (from 300,000-2 million) in the DNA obtained from large cohorts of cases and controls (usually comprising hundreds to thousands of individuals).

Heritability: the proportion of the phenotypic variation in the population, e.g. a disease, eye colour or baldness, that is attributed to genetic variation between individuals.

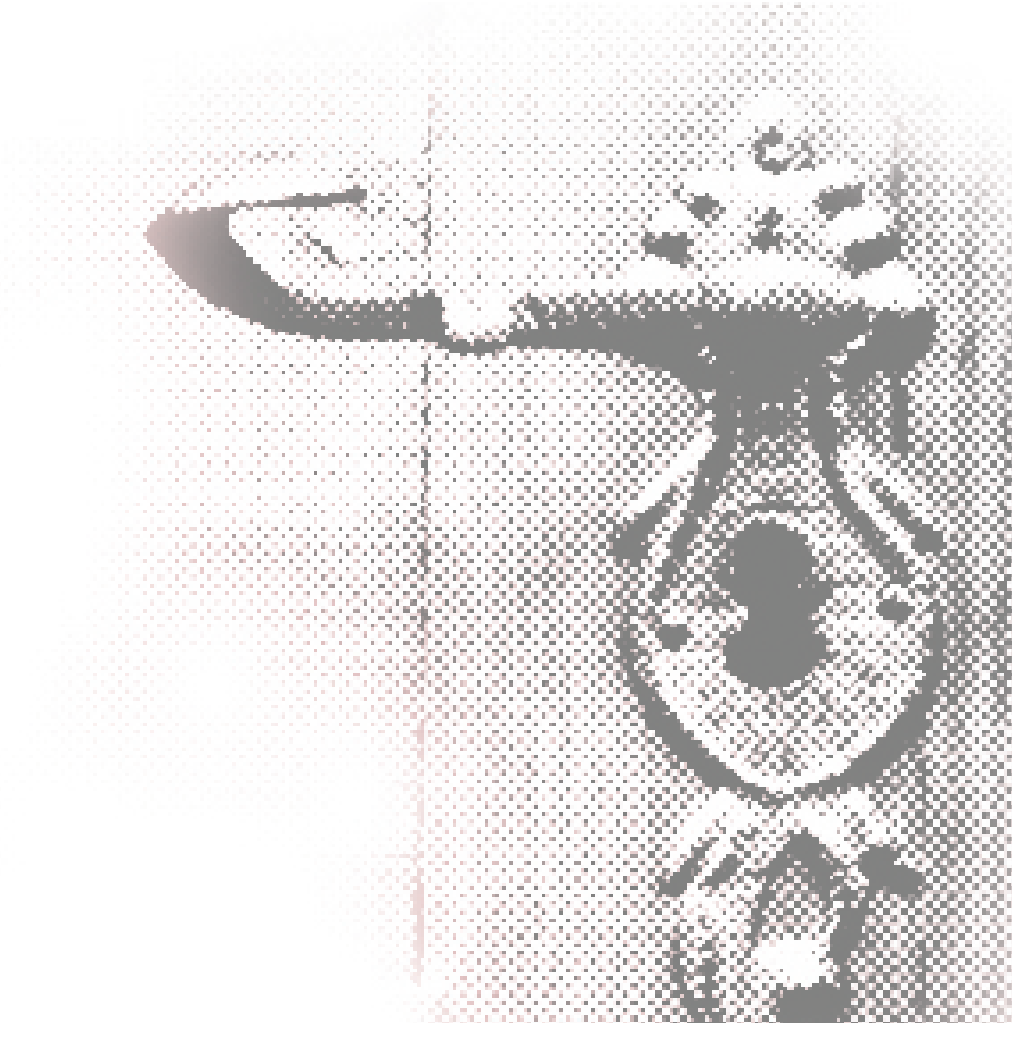
LD: linkage disequilibrium, the correlation between SNPs. The higher the LD, the stronger the correlation is between SNPs.

Locus (plural loci): a region on the DNA.
Sequencing: reading the DNA code.

SNP: a single nucleotide polymorphism, a change in the DNA at a single nucleotide position.



Samenvatting



Genetica is een van het meest dynamische ontwikkelde onderzoeksvelden. Het menselijk genoom is namelijk pas in 2003 volledig gesequenced. Dit was een immense klus die 13 jaar geduurd heeft en $3.000.000.000 gekost heeft. Op dit moment kost het $10,000 om een enkel genoom te sequencen en duurt het korter dan een maand. Door het sequencen van het menselijk genoom weten we dat er ongeveer 10.000.000 veel voorkomende genetische variaties, single nucleotide polymorphisms (SNPs), verandering op een enkele positie in het DNA, bestaan. Het grootste gedeelte van de SNPs zullen een neutraal effect hebben, maar sommige SNPs komen voor met een hogere frequentie in patiënten, dan in gezonde controles. Recente technologische ontwikkelingen hebben geleidt tot platforms die in staat zijn om honderdduizend SNPs in een keer tegelijkertijd te testen (genotyping). Deze platforms worden gebruikt in patiënt-controle (case-control) studies en geven de mogelijkheid om het genotype van iedere SNP te bepalen en om de frequenties tussen patiënten en controles te meten. Statistisch significante verschillen duiden regio's aan in het genoom (disease loci) die belangrijk zijn voor ziekte ontwikkeling. Dit studieontwerp wordt een genoom-wijde associatie studie (GWAS) genoemd. Desondanks, een nadeel van deze aanpak is dat het niet direct het ziekte gen of de mutatie detecteert, maar slechts de regio aanwijst waar het ziekte gen zich bevindt. Deze regio's zijn vaak groot en bevatten vele genen die sterk gecorreleerd met elkaar zijn door de onderliggende linkage disequilibrium (LD) structuur. Mede hierdoor wordt het erg moeilijk om het ziekte gen te detecteren, en moeten we nadat de disease loci geïdentificeerd zijn met GWAS, verdere verfijndere studies uitvoeren om de locatie van het associatie signaal beter te bepalen en de ziekteveroorzakende variant te identificeren.

Door deze technologische ontwikkeling, is er de laatste jaren erg veel inzicht verkregen in de genetische achtergrond van menige ziekten. Dit succes was vooral duidelijk voor monogene aandoeningen, waarbij een verandering in een enkel gen al tot ziekte lijdt. Ongeveer vijf jaar geleden, werd het succes ook merkbaar voor compexe ziekten. Deze aandoeningen ontstaan door kleine, gelijktijdige malfuncties van vele en soms wel honderden genen. Toch is het hier niet alleen maar de genetische component die tot ziekte lijdt, het gaat om het samenspel tussen deze ziekte genen en een of meerdere omgevingsfactoren,

Coeliakie is de meest voorkomende voedsel intolerantie, die bij 1-3% van de westerse populatie voorkomt, en ontstaat door een gluten peptiden intolerantie in genetisch gevoelige personen. Gluten zit in tarwe, rogge, gerst of haver wat veel zit in westerse voedsel producten. Het innemen van gluten veroorzaakt een sterke immuunreactie en veroorzaakt schade aan de dunne darmen. Mensen met coeliakie kunnen leiden aan een heel spectrum van symptomen zoals klassieke klachten diaree, een opgezwollen buik en buikpijn maar ook atypische klachten zoals bloedarmoede en neurologische symptomen.

Genetisch is coeliakie goed gekarakteriseerd door het sterke effect van twee specifieke HLA moleculen, HLA-DQ2 and HLA-DQ8. Deze moleculen verklaren 35-40% van het genetische risico, maar komen ook in 30% van de algemene populatie voor. Dit duidt erop dat andere genetische factoren dus ook betrokken moeten zijn. Sinds 2008, waren er acht niet-HLA loci geïdentificeerd door GWAS in coeliakie patiënten. Het doel van dit proefschrift was het identificeren van andere genen die bijdragen aan de ziekte ontwikkeling.

In hoofdstuk 2, beschrijven we de resultaten van serieuze herhalingsstudies van de coeliakie GWAS. We genotypeerden ongeveer 500 SNPs die gebruikt waren tijdens de GWAS en die een gemiddeld associatie signaal hadden, in 1682 en 3258 controles en vonden 2 extra regio's in het genoom die geassocieerd bleken te zijn. Beide loci bevatten genen die betrokken zijn bij het aansturen van de NF-KB route, een van de essentiële transcriptie factoren die de immuunreactie beïnvloeden.

Immuun- gerelateerde ziekten komen vaak samen voor in families of patiënten, bijvoorbeeld coeliakie patiënten lijden ook vaak aan type 1 diabetes. In hoofdstuk 3 analyseerden wij de genetische overlap van twee auto-immuun ziekten, coeliakie en reuma. We testten of de genen die geassocieerd waren met een van deze ziekten ook bijdragen aan het genetische risico voor de andere. Deze strategie maakte het mogelijk om 6 gedeelde regio's te rapporteren, waarvan 2 loci al eerder gevonden waren en 4 nog niet eerder geïdentificeerd.

Meer coeliakie genen werden ontdekt door een grootschalig GWAS (beschreven in hoofdstuk 4) in 15000 individuen uit vier populaties (UK, Nederland, Italië en Finland). De top 131 SNPs met de sterkste associatie tijdens de analyse in deze populaties werden herhaald in 10000 extra individuen uit de USA, Hongarije, Ierland, Polen, Spanje, Italië en Finland. Met deze aanpak, brachten wij het totale aantal coeliake ziekte loci op 26. Het overgrote deel van deze regio's bevatten genen met een immuun-gerelateerde functie. Behalve genen die betrokken zijn bij het moduleren van de innate en adaptieve immuun respons, vonden we naast genen met een rol in lymfocyt ontwikkeling in de thymus, ook genen betrokken bij de innate immuun detectie van virale infecties. Wat erg belangrijk was dat 50% van de coeliakie geassocieerde genen een sterk effect hadden op gen expressie. Dit suggereert dat veranderingen in gen expressie een van de belangrijkste ziekte veroorzakende mechanismen kan zijn die coeliakie associaties en waarschijnlijk ook andere complexe ziekten onderliggen.

In hoofdstuk 5 geven we een overzicht van de vooruitgang die plaats gevonden heeft door het uitvoeren van genetische studies naar coeliakie en hoe dit er toe bijgedragen heeft in het vergroten van ons inzicht in de moleculaire mechanismen die deze ziekte onderliggen.

Onze Immunochip studie (beschreven in hoofdstuk 6), bevestigde dat disregulatie van gen expressie een van de grootste onderliggende mechanismen is in de ziekte pathogenese. De Immunochip is een op de maat gemaakt platform speciaal ontworpen voor het gebruik in genetische studies naar immuun gerelateerde ziekten. De SNP inhoud is verrijkt voor loci die "immuun genen" bevatten en SNPs die eerder geassocieerd waren met een van de immuun gerelateerde ziekten. De Immunochip was vooral ontworpen om de locatie van het associatie signaal te verfijnen (fine-mappen) in 183 regio's geassocieerd met immuun-gerelateerde ziekten. Voor dit werk, werden deze loci veel verder verrijkt met een hele hoge dichtheid van de zogenoemde genetische markers (SNPs) dan gebruikelijk is in GWAS platforms (10-20x). Een hogere dichtheid van SNPs leidt tot een betere resolutie van het linkage disequilibrium en zorgt voor een betere, preciezere lokalisatie van het associatie signaal. De applicatie van de Immunochip in 12000 coeliakie patiënten en 12000 controles uit zes verschillende landen, zorgde ervoor dat het associatie signaal herleidt kon worden naar een enkel gen in meer dan de helft van de coeliakie loci. Daarnaast, in 20% van de coeliakie loci, was het signaal gelokaliseerd in de regulatoire regio van het gen, wat weer bevestigde dat disregulatie van gen expressie het moleculaire mechanisme kan zijn dat het associatie signaal veroorzaakt. Naast fine-mapping, identificeerde de Immunochip studie 13 andere coeliakie loci, wat het totaal op 39 niet-HLA regio' bracht die coeliakie risico dragen.

Om de genetische architectuur van coeliakie beter te begrijpen, hebben we een zogenoemde cross-etnische analyse gedaan tussen Nederlanders en noord Italianen. We bevestigden en herhaalden 13 van de 26 geteste loci succesvol. We observeerden een verschil in het lange-afstand linkage disequilbrium tussen deze populaties, wat leidde tot een verfijning van de signaal lokalisatie in vijf loci in de noord Italianen en een signaal verschuiving in de andere. Voor een aantal van de loci waren we in staat om kenmerken van positieve selectie druk tegen deze gebieden te rapporteren.

De belangrijkste conclusies van dit proefschrift zijn:

- Tot nu toe zijn 39 niet-HLA loci zijn geassocieerd met coeliakie en samen met HLA verklaren zij 53% van het genetische risico.
- Coeliakie en andere immuun-gerelateerde ziekten delen een groot deel van hun genetische achtergrond.
- Geassocieerde coeliakie loci zijn grotendeels verrijkt met immuun systeem genen.
- Meer dan de helft van de geassocieerde varianten veranderen de expressie van naastgelegen genen.
- Coeliakie geassocieerde varianten zijn meestal niet-coderend en bevinden zich in regulatoire regio's, wat er opduidt dat gen disregulatie een van de belangrijkste moleculaire mechanismen kan zijn voor de pathogenese van de ziekte.
- Dertien coeliakie loci die gelokaliseerd zijn in Europese populaties bevatten ook risico voor coeliakie in een noord Italiaanse populatie.

Vakwoordenlijst

DNA: het molecuul dat alle genetische informatie bevat. De basis eenheid van DNA is een nucleotide.

Genoom: de erfelijke informatie die de ontwikkeling en het functioneren van een levend organisme beschrijft. Het genoom bestaat uit genen en niet-coderende delen (alles wat geen gen is).

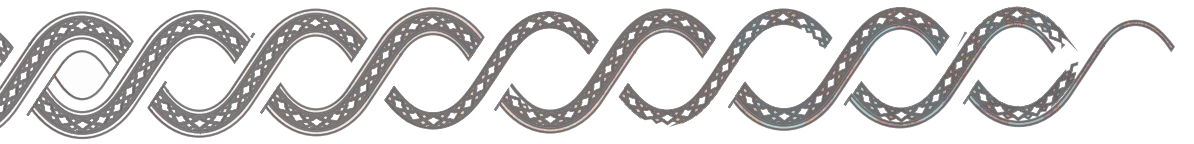
GWAS: Genoom-wijde associatie studie, het genotyperen van honderd duizenden SNPs (300.000-2 miljoen) in het DNA verkregen van grote verzamelingen van patiënten en controles (bestaande uit meestal honderden of duizenden personen).

Erfelijkheid: het gedeelte van de fenotypische variatie in een populatie, dus zoals een ziekte, oogkleur of kaalheid, dat toegekend is aan de genetische variatie tussen individuen.

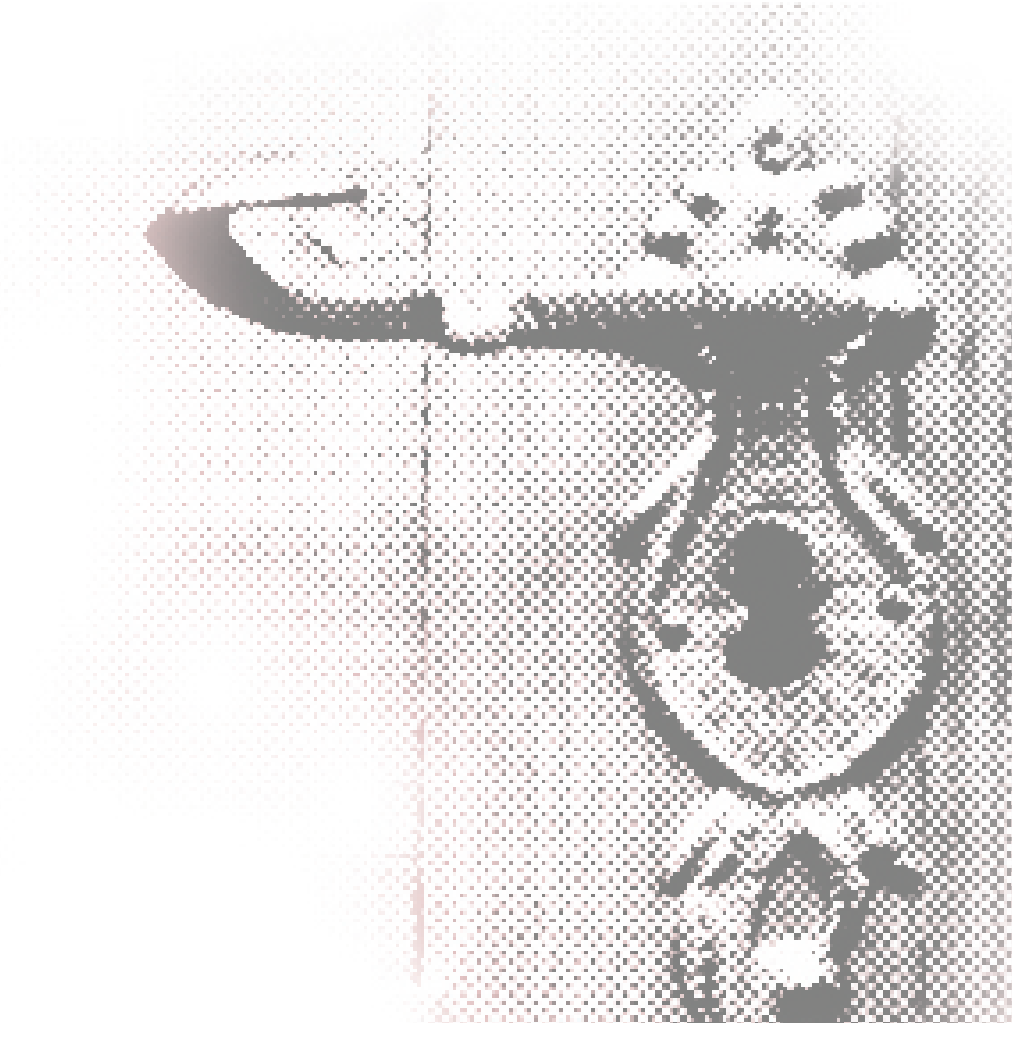
Linkage disequilibruim *(LD)*: de correlatie tussen SNPs. Hoe hoger het LD, hoe sterker de correlatie tussen de SNPs.

Loci (meervoud van locus): een regio op het DNA.
Sequencen: het lezen van de DNA code.

SNP: single nucleotide polymorphism, een verandering in het DNA op een enkele nucleotide positie.



Streszczenie



Genetyka jest jedną z najszybciej rozwijających się dziedzin naukowych. Sekwencjonowanie całego ludzkiego genomu zakończone zostało zaledwie w 2003 roku. Ten przełomowy projekt trwał 13 lat i kosztował trzy miliardy dolarów. Obecnie koszt sekwencjonowania genomu jednej osoby wynosi jedynie 10 tysięcy dolarów i trwa krócej niż miesiąc. Nasz genom zawiera około 10 milionów powszechnie występujących polimorfizmów, zmian w pojedynczym miejscu w DNA (nukleotydzie), (ang. SNP – single nucleotide polymorphism). Większość z tych polimorfizmów jest neutralna i tym samym nie wpływa na rozwój choroby, jednakże niektóre polimorfizmy obserwowane są z większą częstotliwością u chorych w porównaniu do osób zdrowych. W ostatnich latach zaistniał w genetyce ogromny postęp technologiczny, który pozwolił na stworzenie mikromacierzy umożliwiających równoczesne badanie (genotypowanie) setek tysięcy polimorfizmów pojedynczych nukleotydów - SNP (od 300 tysięcy do 2 milionów) u setek, a nawet u tysięcy osób. Mikromacierze te używane są w tzw. badaniach kliniczno-kontrolnych, składających sie z próbek pochodzących od osób chorych i zdrowych, i pozwalają określić genotyp dla każdego polimorfizmu oraz porównać jego częstość występowania pomiędzy chorymi a zdrowymi osobami. SNP, dla których częstość występowania jest statystycznie różna pomiędzy chorymi a zdrowymi wskazują na regiony genomu istotne dla rozwoju choroby. Tego typu test nosi nazwę genomowego badania asocjacyjnego (ang. GWAS – genome-wide association study). Jednakże jedną z wad takich badań jest fakt, że pozwalają one jedynie na identifikację szerokich regionów genomowych, a nie wskazują bezpośrednio na geny lub mutacje prowadzące do rozwoju choroby. Zidentyfikowane regiony często zawierają wiele genów, które są od siebie zależne, zjawisko znane jest jako nierównowaga sprzężeń (ang. linkage disequilibrium – LD). Genomowe badania asocjacyjne rzadko bezpośrednio wskazują na gen odpowiedzialny za wystąpienie choroby, dlatego też po zidentyfikowaniu zasocjowanych regionów chorobowych niezbędne są dalsze badania mające na celu zawężenie obszaru i wskazanie genu lub mutacji prowadzącej do rozwoju choroby (ang. fine-mapping).

To właśnie rozwojowi technologii zawdzięczamy ogromny postęp w rozumieniu podłoża genetycznego wielu chorób. Sukces jest najlepiej widoczny w grupie chorób monogenowych, w których zmiana w pojedynczym genie zawsze prowadzi do rozwoju choroby. Około pięć lat temu sukces zaczął również być widoczny dla chorób o wieloczynnikowej etiologii. Zaburzenia te są rezultatem niewielkich i jednoczesnych dysfunkcji w wielu (nawet kilkuset) genów. Jednak sam komponent genetyczny nie jest wystarczający do wywołania choroby i wiele czynników środowiskowych oddziaływujących z genami w ostateczności prowadzi do jej rozwoju.

Celiakia jest najpowszechniejszą trwałą nietolerancją pokarmową dotykającą około 1-3% zachodnich populacji. Jest wynikiem nietolerancji glutenu u osób predysponowanych genetycznie. Gluten jest białkiem obecnym w pszenicy, jęczmieniu, życie i powszechnie występuje w produktach żywnościowych. Spożycie glutenu powoduje silną, autoimmunologiczną reakcję zapalną i prowadzi do uszkodzenia jelita cienkiego. Chorzy na celiakię mogą prezentować szerokie spektrum objawów, począwszy od klasycznych, jak na przykład biegunka, wzdęcia, niedożywienie oraz bóle brzucha, a skończywszy na atypowych, takich jak anemia, osteoporoza czy objawy neurologiczne.

Celiakia jest dobrze scharakteryzowana genetycznie, od ponad trzydziestu lat wiadomo, że obecność specyficznych genotypów w regionie HLA, dokładnie genów kodujących cząsteczki HLA-DQ2 i HLA-DQ8 jest wysoce skorelowana z chorobą. Genotypy te są odpowiedzialne za 35-40% genetycznego ryzyka wystąpienia choroby, jednakże są one również obecne u ok. 30% osób zdrowych w ogólnej populacji. Fakt ten wskazuje, że do rozwoju choroby niezbędne jest zaangażowanie

innych czynników genetycznych. Do 2008 roku, za pomocą genomowych badań asocjacyjnych, zidentyfikowano osiem obszarów poza regionem HLA, które mogą predysponować do rozwoju celiakii. Celem tej pracy doktorskiej było zidentyfikowanie dodatkowych genów predysponujących do wystąpienia choroby.

W rozdziale drugim przedstawiamy wyniki badań, które opisują genotypowanie około 500 SNP posiadających umiarkowany sygnał asocjacji we wcześniej przeprowadzonym genomowym badaniu asocjacyjnym w celiakii. Polimorfimzy te zostały przetestowane na grupie 1 682 pacjentów z celiakią oraz 3 258 zdrowych osób. To podejście pozwoliło na zidentyfikowanie dwóch dodatkowych regionów w genomie, zawierających geny związane z regulacją aktywności szlaku NF-κB, jednego z kluczowych czynników regulujących odpowiedź immunologiczną.

Choroby immunologiczne często współwystępują z innymi chorobami wśród rodzin lub pacjentów, na przykład chorzy na celiakię często również chorują na cukrzycę typu pierwszego. Rozdział trzeci miał na celu analizę stopnia w jakim celiakia oraz reumatoidalne zapalenie stawów, dzielą między sobą te same czynniki genetyczne. Obydwie choroby mają podłoże autoimmunologiczne. Przetestowaliśmy czy geny predysponujące do jednej choroby stanowią o ryzyku zachorowania na drugą. Ta strategia umożliwiła odkrycie sześciu regionów wspólnych dla tych chorób, potwierdzając dwa poprzednio znane i identfikując cztery nowe.

Kolejne regiony zasocjowane z celiakią zostały zidentyfikowane za pomocą drugiego genomowego badania asocjacyjnego (opisanego w rozdziale czwartym), który został przprowadzony na grupie 15 000 osób pochodzących z czterech różnych europejskich populacji (Brytyjczyków, Holendrów, Włochów i Finów). Aby potwierdzić wyniki z owej analizy, 131 polirmofizmów z najsilniejszym sygnałem asocjacji zostało dodatkowo zgenotypowanych w niezależnej grupie 10 000 osób ze Stanów Zjednoczonych, Węgier, Irlandii, Polski, Hiszpanii, Włoch oraz Finlandii. Ostatecznie liczba regionów zasocjowanych z celiakią wzrosła do 26, z czego wiekszość zawiera geny związane z układem immunologicznym. Poza genami związanymi z regulacją szlaków wrodzonej i nabytej odpowiedzi immunologicznej opisaliśmy również geny zaangażowane w dojrzewanie limfoctów w grasicy oraz geny związane z wykryciem infekcji wirusowych poprzez wrodzoną odpowiedź immunologiczną. Ponadto, wykazaliśmy, że ponad 50% polimorfizmów zasocjowanych z celiakią ma wpływ na ekspresję genów. Sugeruje to, że jednym z głównych mechanizmów prowadzących do rozwoju celiakii, oraz prawdopodobnie innych chorób wieloczynnikowych, będzie zmiana w stopniu ekspresji genów.

W rozdziale piątym omawiamy postęp w badaniach genetycznych nad celiakią oraz zrozumieniu mechanizmów molekularnych będących podłożem choroby.
Badanie z użyciem mikromacierzy „Immunochip" (opisane w rozdziale szóstym) dodatkowo potwierdziło, że głównym mechanizmem powodującym chorobę jest rozlegulowana ekspresja genów. Immunochip został zaprojektowany specyficznie do badań podłoża genetycznego chorób związanych z układem immunologicznym. Polimorfizmy zawarte na tej mikromacierzy reprezentują regiony bogate w geny o funkcji związanej z układem immunologicznym oraz te poprzednio zasocjowane z chorobami immunologicznymi. W szczególności Immunochip został zaprojektowany w celu precyzyjnego zlokalizowania sygnałów asocjacji w 183 regionach związanych z chorobami o podłożu immunologicznym. W tym celu regiony te zostały wzbogacone w dodatkowe markery w rezultacie zwiększając gęstość polimorfizmów od 10 do 20 razy w porównaniu z mikromacierzami używanymi w standardowych genomowych badaniach associacyjnych. Większa gęstość markerów genetycznych zapewnia dokładniejszą analizę korelacji pomiedzy polimorfizmami i pozwala na

bardziej precyzyjne określenie sygnału asocjacji. Zgenotypowanie przy użyciu Immunochip ok. 12 000 pacjentów z celiakią oraz 12 000 osób kontrolnych pochodzących z sześciu różnych krajów pozwoliło na zlokalizowanie sygnału do poziomu pojedynczego genu dla ponad połowy regionów zasocjowanych z celiakią. Dodatkowo, dla 20% z identyfikowanych regionów, sygnał zlokalizowany był w części regulatorowej genu, ponownie wskazując na rozlegulowanie ekspresji genów jako przyczynę molekularnego mechanizmu leżącego u podstaw choroby. Badanie z użyciem Immunochip również zidentyfikowało 13 kolejnych regionów zasocjowanych z celiakią, co daje w sumie 39 regionów, poza HLA, związanych ze zwiększonym ryzykiem zachorowania na celiakię.

Aby dalej zrozumieć genetyczne podłoże celiakii przeprowadziliśmy analizę pomiędzy dwoma etnicznie oddalonymi populacjami. Porównując Holendrów i Hindusów z północy Indii wykazaliśmy, że 13 z 26 poprzednio zasocjowanych regionów również zwiększa ryzyko choroby w populacji północno Indyjskiej. Jednocześnie zaobserwowaliśmy różnicę w korelacji polimorfizmów, w szczególności w długodystansowej nierównowadze sprzężeń, pomiędzy tymi populacjiami. Doprowadziło to do dokładniejszej lokalizacji sygnału asocjacji wśród Hindusów dla pięciu regionów oraz przesunięcia sygnału w pozostałych regionach. Dla kilku regionów również zaobserwowaliśmy działanie pozytywnej selekcji.

Główne wnioski z tej pracy doktorskiej to:

- Poza obszarem HLA do tej pory zostało zidentyfikowanych 39 regionów zwiększających ryzyko zachorowania na celiakię. Wspólnie te 40 regionów wyjaśnia 53% dziedziczności celiakii.
- Celiakia oraz inne choroby związane z układem odpornościowym współdzielą znaczną część podłoża genetycznego.
- Regiony zasocjowane z celiakią w dużej mierze zawierają w geny powiązane z układem odpornościowym.
- Ponad połowa zasocjowanych wariantów wpływa na ekspresję genów.
- Warianty zasocjowane z celiakią są w większości niekodujące oraz zlokalizowane w regulatorowych częściach genomu. Wskazuje to, że rozlegulowana ekspresjia genów może stanowić jeden z głównych, molekularnych mechanizmów prowadzących do rozwoju choroby.
- Spośród regionów zidentyfikowanych w populacjach Europejskich, 13 również stanowi genetyczne ryzyko wystąpienia celiakii w populacji Hindusów z północnych Indii.

Słownik:

DNA: cząsteczka przechowująca informację genetyczną. Podstawową jednostką budującą DNA jest nukleotyd.

Genom: dziedziczna informacja dotycząca rozwoju i funkcjonowania żywych organizmów. Składa się z genów oraz części niekodujących.

GWAS (ang. genome-wide association study): genomowe badanie asocjacyjne polegające na testowaniu setek tysięcy (od 300 tysięcy do 2 milionów) markerów genetycznych, tzw. polimorfizmów (patrz SNP poniżej). Do badań tych wykożystywane jest DNA uzyskane z grup kliniczno-kontrolnych, od osób chorych i zdrowych (zwykle składających się z kilkuset lub tysięcy osób).

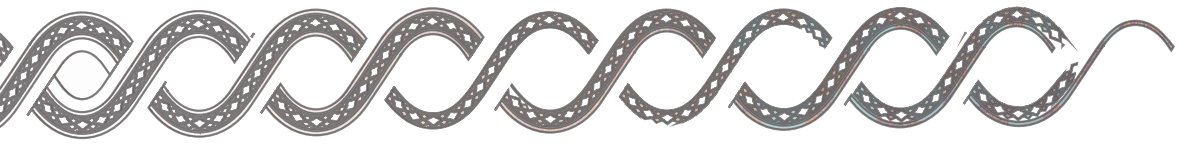
Dziedziczność: część zmienności fenotypowej w populacji, np. choroba, kolor oczu lub łysienie, która przypisywana jest zróżnicowaniu genetycznemu pomiędzy osobami z tej samej populacji.

Nierównowaga sprzężeń (ang. linkage disequilibrium): korelacja pomiędzy polimorfizmami. Im większa nierównowaga sprzężeń, tym większa zależność pomiędzy tymi markerami i prawdopodobieństwo dziedziczenia ich razem.

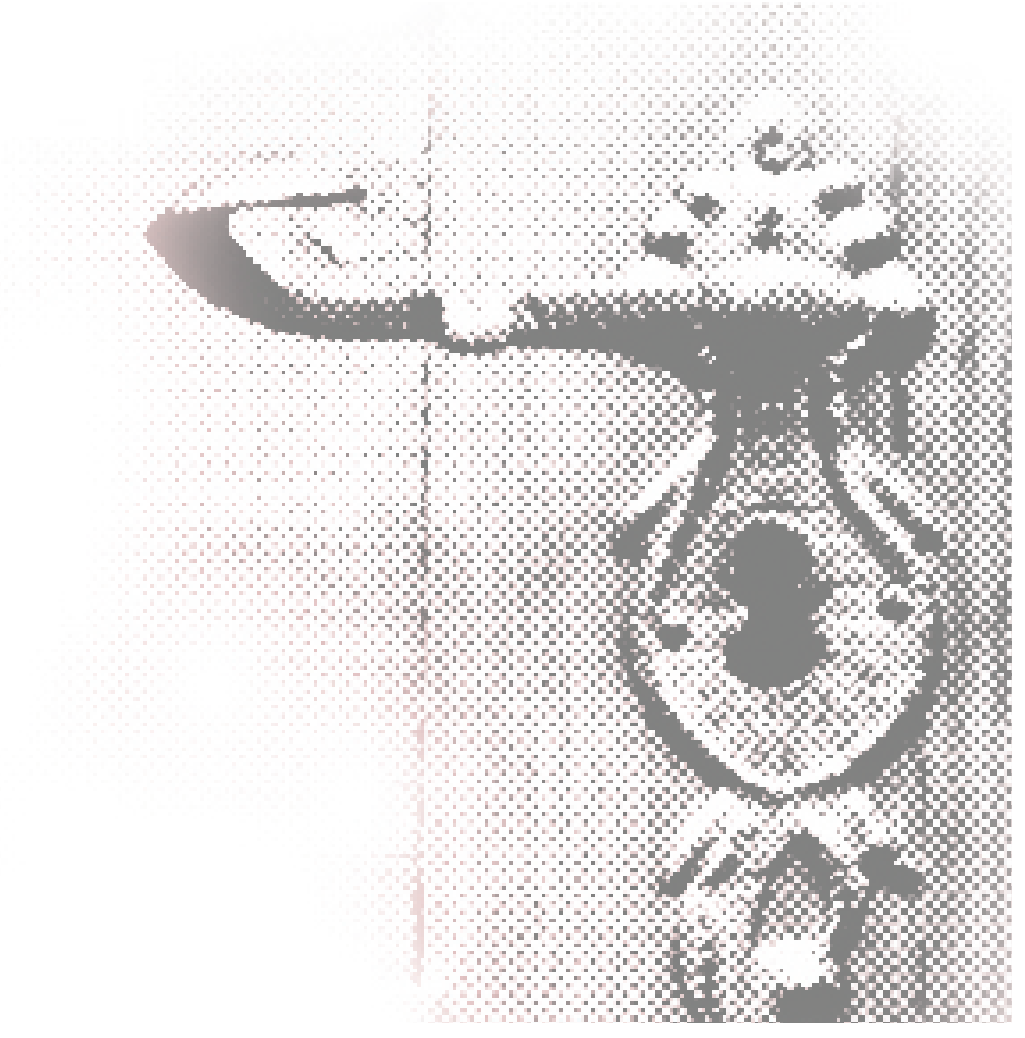
Locus (l.mn loci): region w cząsteczce DNA.

Sekwencjonowanie: odczytywanie kodu DNA.

SNP (ang. single nucleotide polymorphism) – polimorfizm pojedynczego nukleotydu, zmiana w cząsteczce DNA w pojedynczej parze zasad.



Acknowledgements



This 'PhD journey' was way more than just lab work and data analysis. It involved many people who shared the joy, frustration, motivation, ups and downs, and the fun of conducting science. It is inevitable that my journey would never have been so much fun if some of these people had not joined me on this boat. There are way too many of you to thank you all individually here, but I will try to list those who played an essential role in my PhD adventure.

It is clear that my PhD would never have taken shape without the great mentorship from my supervisor. Dear Cisca, thank you for your unbelievable trust in me, I still remember the moment when you told me I would be the one to analyze the GWAS data and I thought that you had completely lost your mind to hand me the responsibility of leading such a massive project (and I mean both the amount of data as well as the amount of money that went through my hands as chips and reagents). Surprisingly (to me but I guess not to you), I managed to get on with this project and it turned out to be a tremendous lesson for me. Not only scientifically, but more importantly it was a lesson that taught me to believe in myself a bit more. I am very grateful to you for all the support I have always received from you, I have learned so much from you. Thank you!

Further, the great support from all my colleagues in the coeliac mafia was essential during these four years.

My SNP adventure started with a person I value very much. Sasha, I admire your optimism, lack of stress (even in the most extreme situations), your scientific criticism and ability to generate great ideas, and most of all I admire you as a person. The way you combine science with your family life is a skill I wish to pursue myself. You are one of the best scientists I have met during these four years. It has always been a great pleasure to brainstorm with you, as well as simply to chat about books or food or go shopping for shoes with a crazy Italian salesman in Montreal. Please don't worry that I will not be in Groningen anymore, the world of science is small and I am sure we will continue to work on some joint projects, simply because it is a lot of fun together!

Dear Jihane, I am extremely happy that we had a chance to be in this PhD boat together! I very much enjoyed our scientific and non-scientific chats! I am not sure how it would all have ended up if not for the support I've had from you in the last years of our PhDs, especially during these last few months of wrapping up our theses. You have the great ability to bring me down to earth, tell me to breath and calm down! Thank you for being a friend and thank you for bringing some fashion into the coeliac group! I wish you lots of success in your new career steps, it will work out great, I am certain about that.

Agata, thank you very much for all the silly jokes and all the laughter we had together in the lab, in the Irish pub and at home. These moments were often a crucial escape from the lab frustrations. It would have been boring if you had not been around. Thank you also for the 'Summary Translation Service', invaluable! *Zawsze będę miło wspominać nasze głupkowate rozmowy. Dziękuję Ci bardzo za wszystko! Trzymam za Ciebie kciuki!…i ryjek do góry ;)*

Lude (although you are not truely the coeliac mafia), thank you for all the great chats we had about science and for all your nasty comments, it was really *COOL* to work with you! Like Cisca, you also shared the madness in believing in my abilities to lead the GWAS. I know you have your opinion on population-specific variants, but I think Cisca and you carry a rare, Dutch-specific haplotype, maybe worth investigating in the future! I look forward to reading your next paper in Science ;)

Isis, Javier, Rodrigo, Juha, Harm-Jan, you guys are like the musketeers to me! I will miss all the chats, parties and BBQs with food and drinks from all over the world, … I wish you all the best! And you'd better stay in touch or I'll come back for your PhD defenses to ask nasty questions! From now on, attending the ASHG is obligatory for you … because of its scientific

value, of course!

Most of my projects could not have been performed at the efficiency and speed that they were carried out without two people, Mathieu and Elvira. You were the greatest companion in the lab! Hybridizinging the GWAS and Immunochip slides together was fantastic fun! I will feel a bit nostalgic when remembering the good, loud, rock music in the Illumina lab, the "whoosh-tips" weapon, silly jokes and comments! My greatest THANK YOU for turning work into great fun!

Asia, thanks for bringing craziness into the group, especially when it comes in bright colours or red trousers! ;) I think it is greatly needed to balance the grey weather outside! Thank you very much for joining forces with Agata in the 'Summary Translation Service'! *Powodzenia ze wszystkim, nie daj się!*

My dear coeliac companions, Astrid, Barbara, Cleo, Jingyuan, Marcel B., Martin W., Ron, Rutger, Sebo, Soesma and Vinod, thank you for sharing your energy and creating positive atmosphere in the group ... even at the crazily early hour of Monday morning work discussions. It was a great pleasure to work with you all. I will miss you! Dear Sabyasachi, thank you for joining us in the group and performing the cross-ethnic study with me. Dasha, I wish you lots of success in your PhD.

Without two other people this thesis would probably never have ended up in time and with all the papers arranged. Dear Jackie, you are the one who always helps, even when it is sent to you at the last minute (which, in my case, unfortunately was most of the time, my apology, one day it will change ... one day). Your input in this thesis is immeasurable! Thank you for all our chats about Britain, Holland, PhD, books ... possibly everything. Thank you for helping me find a place to stay in London. I very much admire you personally with all the positive vibes that you spread around. Dear Hélène, thank you for all your help with all my PhD documents and actually many other issues I had. I don't know how you do it, managing to organize everything, but I am greatly thankful for this magic of yours.

Dear IBD people, Rinse - it was always a big pleasure to see you around. Thank you for our discussions and your very valuable criticism. Noortje, Karin and Suzanne, thank you for your collaboration and support. It was great fun to work with you. I wish you all lots of great publications and success in the IBD genetics.

Genetics brings together biology and bioinformatics, and without the latter it would be very difficult for a biologist to pursue the analysis of genetic data. Therefore a great big thank you to Patrick, Mark-Jan, Roan, Alex, Freerk, Laurent and Morris. Thank you for your help in running all the different analyses I was struggling with. Martijn, thank you for your help on the food database in the gluten-free project.

Similarly, a biologist must acknowledge the help of statistician, dear Gerard, thank you very much for your constructive criticism, very encouraging discussions and enormous patience in explaining the backbone of the haplotype sharing method.

Dear Dineke, Ellen and Robert, thank you for our conversations and discussions. I wish you and your teams success in obtaining break-through scientific findings and occupying many pages of Nature, Cell and Science. Dineke, a big thank you for helping with the translation of my Dutch summary.

Dear Paul, it was a great pleasure to be your office-mate. You bring in a load of good energy, laughs and nice chats. Thanks for that! And speaking of talking, dear Mats, thank you for your long, philosophical discussions. I am very happy we shared the same office during our PhD. I am happy to have not only a colleague but also a friend in you. I wish you lots of success in science as well as in your private life. Please send me a copy of your next Cell paper!

Dear Bote, Mentje,Ria and Joke, thank you all for your help in arranging all sorts of documents I came to you with.

Bahram and Mariska, thank you for your help and the good atmosphere in the genotyping lab. Pieter, thank you for your assistance with managing the coeliac Illumina orders.

My dear lab companions, Helga, Yunia, Gerben, Jana, Mateusz, Annemieke, Cristine, Marcel, Anna Duarri, Rian, Justyna, Maria, Ana Fereira, Anna Pósafalvi, Omid, Céline, Rajendra and Olga, thank you very much for creating a good atmosphere in the lab. I thank all my colleagues in the lab, technicians, students and others, who created the great atmosphere at work.

My journey with SNPs and coeliac disease started in Utrecht, where a group of great people helped me find my way through the subject of complex genetics. Dear Bobby, Roel, Erik, Clara, Carolien, Dalila, Harm, Frederike, Karen, Ruben, Behrooz and Linda, thank you very much for your warm welcome in the lab there. Your discussions and scientific criticism definitely inspired me to pursue my PhD in complex genetics. Maciek, thank you for showing me the route from the Biltstraat to the lab on the first day of my work in Utrecht. *Dziekuję za cierpliwość by tłumaczyć mi genetykę na początkach mojego doktoratu.*

As is clearly reflected by the number of co-authors that contributed to the papers in this thesis, my work would have not been possible without the collaborators from all over the world.

Dear David, thank you for taking me into your group to jointly analyze the GWAS and Immunochip data. I enjoyed my time in your group very much and have learned a lot. I wish you all the best and hope we have a chance to collaborate again. Dear Karen and Patrick, it was a great experience to perform these two big projects jointly. I have enjoyed working with you very much. Graham, Nick and Vanisha, thanks a lot for all the chats and curry lunches together. Nici, thank you for your help with arrangements regarding my stay at ICMS.

Vincent, thank you very much for all your help with the statistics in the Immunochip study.

Jeff, thank you for your amazingly quick replies to all my bothersome e-mails and your help in the GWAS and Immunochip studies.

During my visit to London I came across many people who made my stay great fun. Dear Mahmood, James, Neil, Manoj and Naheed, thanks for funny chats at ICMS, during curry lunches and in the pub.

Dear Ross, Donatella, Luigi, Bożena, Thelma, Maria Cristina, Jose, Paivi, Susan, Carlo, Mihai, Santos and Marieke C, thank you for your collaborations and feedback on our joint projects. I wish you lots of success and many great publications.

Members of the CDC consortium – thanks for your collaboration and discussions. I am, of course, also grateful to all the patients who participated in the research, and to the Nederlands Coeliakie Vereniging for their help. Chris, Luisa, Roderick, Victorien, Sjoerd, Marko and Geertje – thank you for help with phenotype classifications and for the excellent collaboration in the coeliac research.

Dear Marten, it was always a pleasure to have discussions with you.

Dear Robert, Soumya and Elia, thank you for your collaboration on our joint projects. Dear Dorothée, thank you for organizing a great dinner in Montreal. I am very much looking forward to joining you on board in Boston.

Claudia, thank you very much for all your help in putting my thesis together, into this great layout.

Aleku, dziękuje za dach nad głową gdy musiałam lecieć z Amsterdamu o jakiejś barbarzyńskiej porze. W połaczeniu z twoimi kulinarnymi

umiejętnościami i dobrą wódka był to najlepszy z możliwych noclegów. Powodzenia z doktoratem, to już koncówka, później musi być lepiej!

Agnieszko i Wojteku, dziękuję bardzo za waszą gościnę i dach nad głową w Cambridge. Chętnię sie odwdzięczę i zapraszam do Bostonu.

Magier paskudo, dziękuje za wszystkie durne wymiany zdań. Trzymam kciuki za Twój dokotorat.

Musiałku i Zdebie, moi drodzy, dziękuję wam bardzo za wsparcie z Krakowa, zwłaszcza to intensywne przy moich wizytach w Polsce ;)

I was very lucky to make a lot of friends here in Holland. Dear all, thank you for your openness and support.

My dearest Tina ... pigeon ... trrrrrr! You are the best Greek on this planet! I am your single person fun club! I feel honoured to have such a silly best friend as you! Our time at the Biltstraat, the skype chats and travels we had during our PhDs will remain one of my best memories! Thank you for being there for me and for your Greek charm that lightened up my days in Holland! Hope we end up in Boston togehter. Remember, the wind is always against you! Can you dance ... *merengue*?!

Monique, thank you for being one of the best friends I could ask for. Thank you for all the support you gave during my stay in Groningen, the long chats we had, all the Guinness pints in O'Ceallaigh's, your visits to London, our festivals, cycling, and most of all the tears and laughter we had together. You are my sparkle in Groningen, please never change.

Dear Ania, you are another sparkle of mine. Your optimism, loud laugh and the tones of positive energy that you glow with are irreplaceable! Your dedication for patients and combining an MD with science, even though it's so hard in Poland, is truly admirable. *Rybaku mój drogi, podziwiam Cię bardzo za Twoje ambicje, serce do walki o realizowanie Twoich pasji, ogrom energii i niezliczone pokłady empatii jakie masz w sobie.*

My dear Mitja, I am very glad that, despite your attachment to Southern Europe, you undertook the trip to the far north of the Netherlands and ended up in Groningen. Thank you for the support you give me (especially during the last few months) and for showing me the beauty of your beloved Balkans. *Želim si, da bi na najini barki jadrala še dolgo časa.*

Last but not least, the most important acknowledgements go to my parents, grantparents, brothers and the whole family. My dearest ones, without your support it would have been a tough road. I thank you from all my heart for all your warm words and trust in me.

Mom and dad, without you I would not be here now. I know it is not easy to have three children and see them spread all over the world for most of the time, therefore even more I want to thank you for your support and believing in me.

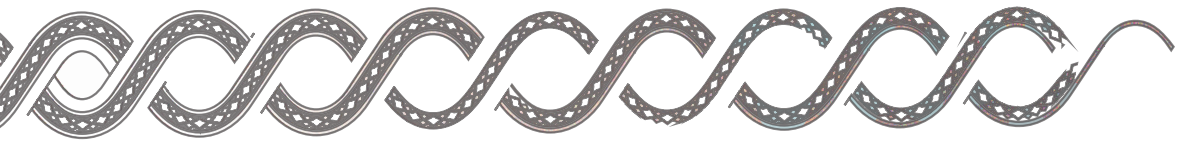
Najważniejsze podziękowania kieruje swoim rodzicom, dziadkom, braciom i całej rodzinie. Moi drodzy, bez waszego wsparcia byłaby to bardzo trudna droga do przebycia. Z całego serca dziękuję wam za ciepłe słowa i wiarę we mnie.

Moja droga mamo i babciu Pelagio, jesteście najważniejszymi kobietami, w moim życiu. Wy nauczyłyście mnie jak być ciekawą świata, otwartą na ludzi i jednocześnie umieć walczyć o siebie. Jestem wam niezmiernie wdzięczna za tę lekcję. Bardzo was podziwiam.

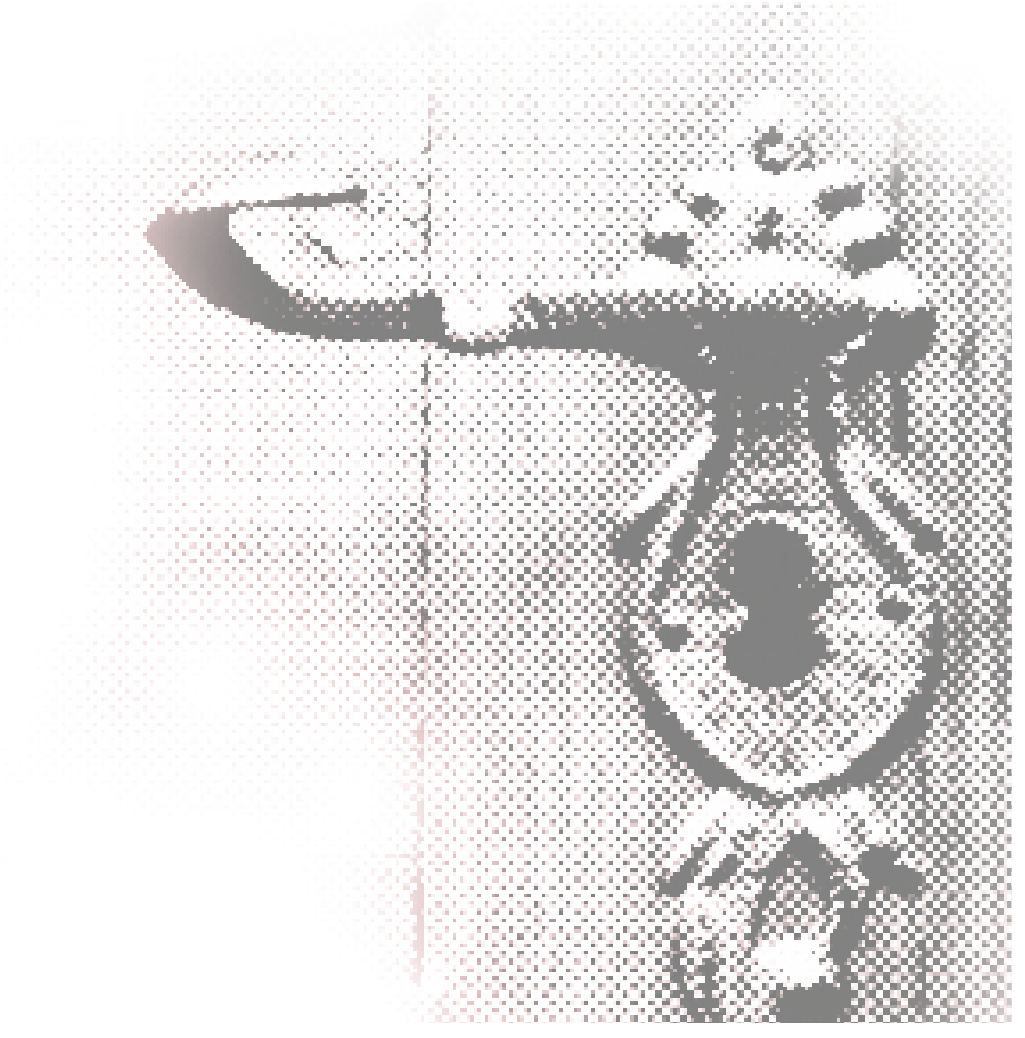
Mamo i tato, gdyby nie wy, nie byłoby mnie tutaj. Wiem, że nie jest łatwo mieć trójke dzieci, które większość czasu są rozproszone po świecie, dlatego tym bardziej dziękuję wam za wasze wsparcie i wiare we mnie przez wszystkie lata. Kocham was bardzo. Wam dedykuję tę pracę.

Thank you ... *bedankt ... dziękuję!*

Gosia



Curriculum vitae



Małgorzata, Barbara (Gosia) Trynka was born on 13th July 1983 in Kraków, Poland. In 2002 she started an MSc course in Biotechnology at the Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Kraków, Poland. In 2006-2007 she worked on a collaborative project with Prof. Cisca Wijmenga at the Department of Medical Genetics at Utrecht University, the Netherlands; this aimed to identify sequencing and splicing variants in the MYO9B gene in relation to predisposition to coeliac disease. In August 2007, after graduating with an MSc degree, she started her PhD research at the Department of Genetics, University Medical Centre Groningen, working on unravelling the genetics of coeliac disease.

During her PhD work, she was awarded a *Ter Meulen Fund* grant from the Royal Netherlands Academy of Arts and Sciences (KNAW) for a 4-month visit to the group of Prof. Van Heel at the Blizard Institute of Cell and Molecular Science, Barts and the London, Queen Mary's School of Medicine and Dentistry, London, UK. In addition, she was awarded many scholarships and prizes, including the *Lodewijk Sandkuijl* prize for an outstanding presentation in the field of complex genetics and statistical genetics at the European Human Genetics Conference (Gothenburg, Sweden, 2010); was a finalist for the Trainee Research Award for an abstract submitted to the 12th International Congress of Human Genetics (Montreal, Canada, 2011) and received several *Simonsfonds* travel grants.

She has been invited as a guest lecturer in the Netherlands and at a conference on Genetics in Gastrointestinal and Liver Diseases in Romania. She also gave a 1-week course at the University of Delhi, India.

In 2012 she will take up a postdoctoral fellowship with Dr. Robert Plenge and Dr. Soumya Raychaudhuri in the Department of Rheumatology, Harvard Medical School, Boston, USA.

List of publications

2011

Trynka G#, Hunt KA#, Bockett NA, Mistry V, Bakker SF, Bardella MT, Barisani D, Bhaw-Rosun L, Bilbao JR, BK T, de la Concha EG, Cukrowska B, Dias KRM, Dubois PCA, Edkins S, Franke L, Fransen K, Greco L, Gutierrez J, Heap GAR, Hunt S, Izurieta LP, Joosten L, Langford C, Mazzilli MC, Mearin ML, Mein CA, Midah V, Mitrovic M, Mora B, Nutland S, Núñez C, Onengut-Gumuscu S, Pearce K, Platteel M, Polanco I, Potter S, Rich SS, Romanos J, Rybak A, Santiago JL, Senapati S, Sood A, Sperandeo MP, Szperl A, Varadé J, Wolters VM, Zhernakova A, PreventCD Study Group, Wellcome Trust Case Control Consortium, Urcelay E, Duerr RH, Plagnol V, Barrett JC, Deloukas P, Wijmenga C, van Heel DA. (2011). Dense genotyping reveals and localises multiple common and rare variant association signals in celiac disease. *Nature Genetics in press*

Fehrmann RSN, Jansen RC, Veldink JH, Westra H, Arends D, Bonder MJ, Fu J, Deelen P, Groen HJM, Smolonska A, Weersma RK, Hofstra RMW, Buurman WA, Rensen S, Wolfs MGM, Platteel M, Zhernakova A, Elbers CC, Festen EM, **Trynka G**, Hofker MH, Saris CGJ, Ophoff RA, van den Berg LH, van Heel DA, Wijmenga C, te Meerman GJ, Franke L. (2011). Trans-eQTLs reveal that rndependent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *Plos Genetics*; 7:e1002197.

Sperandeo MP, Tosco A, Izzo V, Tucci F, Troncone R, Auricchio R, Romanos J, **Trynka G**, Auricchio S, Jabri B, Greco L. (2011) Potential celiac patients: a model of celiac disease pathogenesis. *PLoS One*;6:e21281.

Szperl A, Ricaño-Ponce I, Li J, Deelen P, Kanterakis A, Plagnol V, van Dijk F, Westra H, **Trynka G**, Mulder C, Swertz M, Wijmenga C, Zheng Hch. (2011). Exome sequencing in a family segregating for celiac disease. *Clin ical Genetics.*;80:138-47.

Zhernakova A, Stahl EA, **Trynka G**, Raychaudhuri S, Festen EAM, Kurreeman F, Franke L, Fehrmann RSN, Thomson B, Gupta N, Patsopoulos N, Romanos J, McManus R, Ryan AW, Turner G, Remmers EF, Greco L, Toes R, Grandone E, Mazzilli MC, Rybak A, Cukrowska B, LiY, de Bakker P, Gregersen PK, Worthington J, Siminovitch K, Klareskog L, Huizinga T, Wijmenga C, Plenge RM. (2011). GWAS meta-analysis of celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *Plos Genetics*;7:e1002004.

Festen EAM, Goyette P, Green T, Beauchamp C, Boucher G, **Trynka G**, Dubois PC, Stokkers PCF, The International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), Hommes DW, Barisani D, Palmieri O, Annese V, van Heel DA, Weersma RK, Daly MJ, Wijmenga C, Rioux JD. (2011) A meta-analysis of genome-wide association scans identifies TAGAP and PUS10 as shared risk loci for Crohn's disease and celiac disease. *Plos Genetics*; 27;7;e1001283.

2010

Trynka G, Wijmenga C, Van Heel DA. (2010) A genetic perspective on coeliac disease. *Trends in Molecular Medicine*; 16:537-50.

Zhernakova A, Elbers CC, Ferwerda B, Romanos J, **Trynka G**, Dubois PC, de Kovel CG, Franke L, Oosting M, Barisani D, Bardella MT; Finnish Celiac Disease Study Group, Joosten LA, Saavalainen P, van Heel DA, Catassi C, Netea MG, Wijmenga C. (2010) Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *American Journal of Human Genetics*; 86:970-7;

Trynka G#, Dubois PC#, Franke L, Hunt

KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adány R, Aromaa A, Bardella MT, van den Berg LH, Bockett NA, de la Concha EG, Dema B, Fehrmann RS, Fernández-Arquero M, Fiatal S, Grandone E, Green PM, Groen HJ, Gwilliam R, Houwen RH, Hunt SE, Kaukinen K, Kelleher D, Korponay-Szabo I, Kurppa K, MacMathuna P, Mäki M, Mazzilli MC, McCann OT, Mearin ML, Mein CA, Mirza MM, Mistry V, Mora B, Morley KI, Mulder CJ, Murray JA, Núñez C, Oosterom E, Ophoff RA, Polanco I, Peltonen L, Platteel M, Rybak A, Salomaa V, Schweizer JJ, Sperandeo MP, Tack GJ, Turner G, Veldink JH, Verbeek WH, Weersma RK, Wolters VM, Urcelay E, Cukrowska B, Greco L, Neuhausen SL, McManus R, Barisani D, Deloukas P, Barrett JC, Saavalainen P, Wijmenga C, van Heel DA. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*; 42:295-302.

2009

Trynka G#, Coenen MJ#, Heskamp S, Franke B, van Diemen CC, Smolonska J, van Leeuwen M, Brouwer E, Boezen MH, Postma DS, Platteel M, Zanen P, Lammers JW, Groen HJ, Mali WP, Mulder CJ, Tack GJ, Verbeek WH, Wolters VM, Houwen RH, Mearin ML, van Heel DA, Radstake TR, van Riel PL, Wijmenga C, Barrera P, Zhernakova A. (2009). Common and different genetic background for rheumatoid arthritis and coeliac disease. *Human Molecular Genetics*; 1;18:4195-203.

Romanos J, van Diemen CC, Nolte IM, **Trynka G**, Zhernakova A, Fu J, Bardella MT, Barisani D, McManus R, van Heel DA, Wijmenga C. (2009) Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. *Gastroenterology*; 137:834-40, 840.

Trynka G#, Zhernakova A#., Romanos J, Franke L, Hunt K, Turner G, Platteel, M, Ryan, AW, de Kovel C, Barisani D, Bardella MT, McManus R, Van Heel DA, Wijmenga C. (2009) Coeliac disease associated risk variants in TNFAIP3 and REL implicate altered NF-κB signalling. *Gut*; 58:1078-83.

Elbers CC, de Kovel CG, van der Schouw YT, Meijboom JR, Bauer F, Grobbee DE, **Trynka G**, van Vliet-Ostaptchouk JV, Wijmenga C, Onland-Moret NC. (2009) Variants in neuropeptide Y receptor 1 and 5 are associated with nutrient-specific food intake and are under recent selection in Europeans. *PLoS One*, 17;4:e7070.

Romanos J, Barisani D#, **Trynka G#**, Zhernakova A, Bardella MT, Wijmenga C. (2009). Six new celiac disease loci replicated in an Italian population confirm association to celiac disease. *Journal of Medical Genetics*. Jan;46:60-3.

2008

Heap GA, **Trynka G#**, Jansen RC#, Bruinenberg M, Dinesen LC, Hunt KA, Wijmenga C, Van Heel DA & Franke L. (2008). Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Medical Genomics*; Jan 7;2:1.

Franke LH, de Kovel CGF, Aulchenko YS, **Trynka G**, Zhernakova, A.P., Hunt KA, Blauw HM, van den Berg LH, Ophoff RA, Deloukas P, Van Heel DA, Wijmenga C. (2008). Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *American Journal of Human Genetics*, 82(6), 1316-1333.

Hunt KA, Zhernakova AP, Turner G, Heap GA, Franke LH, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle P, Walters JR, Sanders DS, Playford RJ, **Trynka G**, Mulder CJ, Mearin ML,, Verbeek WH, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle W, Mein CA, Wapenaar MC, Deloukas P,

McGinnis R, McManus R, Wijmenga C, Van Heel DA. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genetics*, 40(4), 395-402.

Zhernakova AP, Festen EM, Franke LH, **Trynka G**, van Diemen CC, Monsuur AJ, Bevova MR, Nijmeijer RM, van ´t Slot R, Heijmans R, Boezen HM, Van Heel DA, van Bodegraven AA, Stokkers PC, Wijmenga C, Crusius JBA, Weersma RK. (2008). Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. *American Journal of Human Genetics*, 82(5), 1202-1210.

#equal contribution